Identification and Search of Protein Structural Motif

Yu-Chiang Kuo

NCLab Report No. NCL-TR-2009006
July 2009

Natural Computing Laboratory (NCLab)
Department of Computer Science
National Chiao Tung University
329 Engineering Building C
1001 Ta Hsueh Road
HsinChu City 300, TAIWAN
http://nclab.tw/

國立交通大學

資訊科學與工程研究所

碩 士 論 文

蛋白質結構模板之識別與搜尋

Identification and Search of Protein Structural Motif

研 究 生：郭育強

指導教授：陳穎平　教授

中 華 民 國　九 十 八 年 七 月

蛋白質結構模板之識別與搜尋
Identification and Search of Protein Structural Motif

研 究 生：郭育強　　Student：Yu-Chiang Kuo

指導教授：陳穎平　　Advisor：Ying-ping Chen

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

July 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年七月

蛋白質結構模板之識別與搜尋

學生：郭育強　　指導教授：陳穎平

國立交通大學資訊科學與工程研究所碩士班

摘　　要

在蛋白質科學領域裡，一般的認知是一胺基酸序列決定蛋白質結構，蛋白質結構決定蛋白質功能。隨著蛋白質結構數量的快速成長以及蛋白質體學技術的演進，許多研究胺基酸序列、蛋白質結構以及蛋白質功能之間關係的方法被不斷地提出。

本論文的研究是藉由檢討 PROSITE 資料庫來探討胺基酸序列、蛋白質結構以及蛋白質功能三者之間的關係。PROSITE 是一個被廣泛應用，儲存完整的生物模板及其功能註解的資料庫。我們檢討 PROSITE 蛋白質模板在結構面的保守性，藉以驗證蛋白質結構會導引蛋白質功能的基本信條。

我們發展了一套新的工具「fastCOPS」，其邏輯流程整合 3D-BLAST 做為快速搜尋、MAMMOTH 做為精確結構比對方法，以及遞迴截取機制。一般來說，只要現行工具及方法能夠相容於 fastCOPS 的設計，就能夠套用為架構元件。做為快速搜尋的元件，必須能夠接受一個蛋白質片段作為輸入；做為精確結構比對元件，必須具有比對部份結構並可適當插入間隔的功能。

我們應用了 fastCOPS 做為蛋白質結構模板搜尋以及識別的工具。fastCOPS 利用許多的蛋白質結構模板做測試，包括 treble clef finger 以及 leucine-rich repeat。另外，我們亦利用了 PROSITE 的蛋白質模板作為 fastCOPS 輸入，來展示 fastCOPS 能夠搜尋到在蛋白質結構方面具有保守性的相似片段，而若利用胺基酸序列搜尋方法卻是難以達成的能力。

關鍵詞：fastCOPS、BLAST、結構比對、區域保守性、結構模板搜尋

Abstract

In protein science, the common belief is that amino acid sequence determines protein structure, and then protein structure determines the biological function. As the availability of the rapidly growing number of protein crystal structures and the advent of proteomics technologies, many methods have been proposed to identify sequence-structure-function relationships.

In this study, we investigate the relationship of protein sequence-structure-function by surveying PROSITE database. PROSITE is a widely used database that maintains annotated biological motifs. We review the structurally conserved property of PROSITE patterns to validate the fundamental principle—protein structure leads to protein function.

In addition, we proposed fastCOPS that integrates a quick screening method, 3D-BLAST, an accurate structural alignment method, MAMMOTH, and the mechanism of recursive truncation with an appropriate logic flow. In general, tools and methods currently available can be adopted in the fastCOPS framework as long as they are compatible with the design. The quick component should be able to accept a protein fragment as input, and the accurate component has to be capable of aligning partial structures with possible gap insertions.

We apply fastCOPS to achieve the task of structural motif search and identification. The fastCOPS has been evaluated on various structural motifs, including the treble clef finger motif, and the leucine-rich repeat motif. In addition, we use a PROSITE pattern as query to demonstrate the capability that fastCOPS can find structurally conserved fragments but using sequence alignment tool will hardly be achieved.

keywords:

fastCOPS, BLAST, structural alignment, local conserved, structural motif search

Acknowledgements

I would like to express my most sincere gratitude to my advisor, Dr. Chen, for everything he does that conducts me to learn what a solid research is. I am very thankful to learn with him in these two graduate years.

I would like to convey my many thanks to the committee members, Dr. Horng and Dr. Yang, for giving me so valuable opinions and suggestions that make my thesis more complete.

I would like to express my many thanks to Dr. Tung, who assists me a lot to overcome difficulties in my thesis. I would also thank my Natural Computing Laboratory members, Jih-Yiing and Pei Jiang. They are so friendly and obliging.

I would like to appreciate my parents, who always give me the best support and care in my life.

Finally, I would like to offer my thanksgiving to God for all the wonderful things he has done.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In protein science, the common belief is that amino acid sequence determines protein structure, and then protein structure determines the biological function [1]. As the availability of the rapidly growing number of protein crystal structures and the advent of proteomics technologies, many methods have been proposed to identify sequence-structure-function relationships.

In order to realize the biological function of new proteins, applying sequence alignment or comparing their folds with all known structures is usually prior approach. However, sequence-based methods and fold comparison do not always provide sufficient information to predict function. In this situation, finding local conserved structure (structural motif) is a feasible method. The structural motifs are composed of a few secondary structure and often have functional significance. They can be treated as minimal functional unit in a protein [2].

Structural motif is fundamental for many applications, such as protein structure prediction, fragment library design, antibody design, and drug design. For identifying structural motifs to annotating the functions of a newly determined structure, pair alignments and database search methods play key roles.

However, detailed protein structure alignment methods often provides a barely satisfactory response time for large databases with tens of thousands of structures. The promise of the protein structure databases continuous grow demands further improvement in terms of the computational efficiency of structural database search methods.

1.2 Related Works

The most widely used and famous sequence alignment method is BLAST [3]. BLAST is a efficient tool which enable researchers to rapidly scan the entire database with a complete or partial sequence as query and deliver accurate sequence alignment results with statistical significance. However, due to evolutionary distance increasing, the homologous sequence similarity gradually declined. Sequence alignment algorithm often could not provide detecting distantly related proteins in homologous relationships.

Owing to protein structures have conserved property better than amino acid sequences, the information of biological function or the evolutionary relation of proteins can be supplied by structural comparison [4].

Many structure comparison methods have been developed, such as CE [5], DALI [6], and VAST [7]. The combinational extension (CE) algorithm uses octameric fragment pairs for aligning between two structures. When significant alignment occurs, it uses dynamic programming approach to perform further optimal alignment repeatedly [5]. DALI, presents protein structures as 2D distance matrices between C_α atoms to be compared and use Monte Carlo method to optimize the matrices [6]. The vector alignment search tool (VAST), describes protein secondary structure elements (SSEs) as vectors to derive the topology of the structure and base on Gibbs sampling to perform optimal alignment [7].

These methods compare a pair of known structures and deliver accurate structure alignment results with statistical significance. However, they are not appropriate to be applied on databases scanning task.

In order to reduce the computational overhead while scanning huge structure databases, using structural alphabet to encode 3D space information as 1D sequence and utilize some sequence alignment methods for structure alignment is feasible approach. 3D-BLAST [8] use nearest-neighbor algorithm to cluster structural segments. They encoded structural segments to 23 letters according to their features of kappa and alpha angle. Through their customized substitution matrix for structural alignment, based on the advantage of BLAST, 3D-BLAST successfully demonstrated efficient performance on scanning structure databases. SA-FAST [9] uses artificial neural network method–SOM (self-organizing

map) and minimum spanning tree algorithm to determine the structural alphabet size. Then, they use k-means algorithm to cluster protein structure segments, and use the centroid of the clusters to define their structural alphabet. They also customize specific substitution matrix for applying their structure alphabet on traditional sequence alignment tool–FASTA.

The developed methods belong to the category of local structure search, such as FF (Fragment Finder) [10] and PAST [11]. FF compares protein structures based on main chain backbone conformational phi and psi angles, and allow users define structural fragment they interested to search similar structure [10]. PAST pre-processes the protein database and create a specific data structure–suffix tree containing the backbone information to speed up the structure search [11]. These methods increase the flexibility of structure comparison significantly. However, local conserved structures (structural motif) often occur repeatedly and even overlapping. How to raise the sensitivity in this kind case seems be neglected. Another question, short fragment searching may cause high false positive, so how to select an appropriate fragment as query for avoiding false positive is also a issue needs to discuss.

1.3 Thesis Objectives

In this thesis, we investigate the relationship of protein sequence-structure-function by surveying PROSITE database. Figure 1.1 shows how we survey the PROSITE patterns through structural alignments and review the PROSITE PDOC to investigate the relationship of structure and function. PROSITE is a widely used database that maintains annotated biological motifs. We review the structurally conserved property of PROSITE patterns to validate the fundamental principle—protein structure leads to protein function [2].

Furthermore, we propose a framework called "fastCOPS" that can rapidly identify local conserved structure by searching protein structure databases. Equipped with the mechanism of recursive truncation, the fastCOPS enables to search the entire Protein Data Bank (PDB) [12] for similar local, conserved structure of a query structure.

1.4 Road Map

- Chapter 1 consists of the motivation, related works, objectives and organizations of this study. It describes why this research is important and the main tasks to be accomplished.

- Chapter 2 describes that we collect structural fragments corresponding the sequence region described by PROSITE pattern entries to proceeding structural alignment using the popular structural alignment tool–MAMMOTH. Then, for judging structure similar degree, we calculate the RMSD (Root Mean Square Deviation) value within the same pattern entries as intra-pattern alignment and distinct pattern entries as inter-pattern alignment. In order to verify the two distributions (intra and inter) are distinct, we use the non-parametric statistic method–Wilcoxon rank-sum test. Through the test, we evaluate the two distributions are significantly distinct.

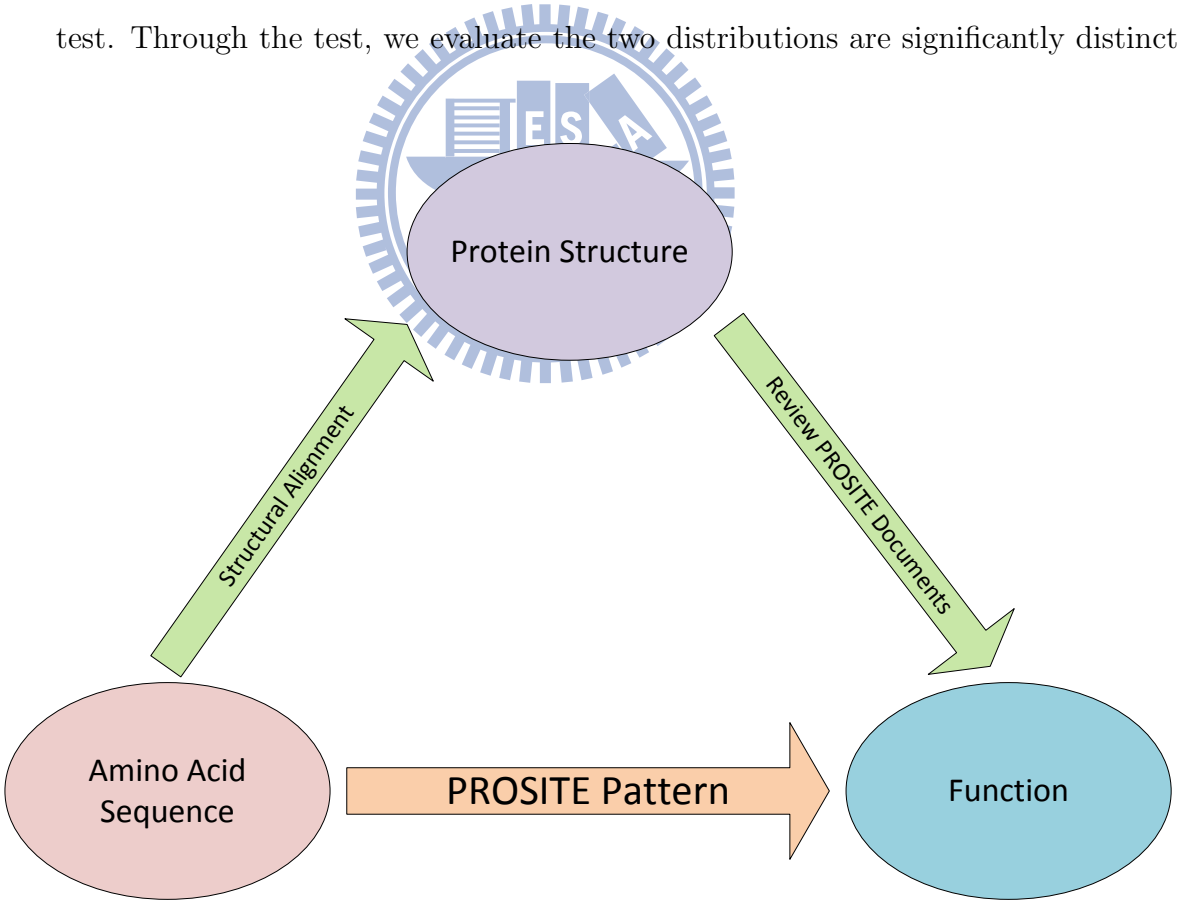

Figure 1.1: Through pairwise structural alignments between PROSITE patterns, we intend to confirm the structurally conserved properties within the patterns. Then, by reviewing the PROSITE PDOC, we investigate whether structurally conserved fragments reflect similar biological function.

So, we focus on exploring the relationship of structure similarity and functional identity.

We design a re-clustering procedure to group PROSITE patterns with similar conformation. Through observing the results of re-clustering, we can review the PROSITE documents to identify whether similar structures reflect functional identity. The observation motivate us to develop a structure searching tool.

By the filter-and-refine framework, we integrate 3D-BLAST, MAMMOTH and embed the recursive truncation procedure to develop a local conserved structure searching tool, i.e., fastCOPS.

- Chapter 3 presents the results of structural alignment and interpret the results after re-clustering procedure executed. According to the results of re-clustering, we list several cases to be illustrations explaining the relationship of similar structure and functional identity. However, we also observed some strange cases that grouped too many PROSITE patterns. In order to discover the over-grouping reason, we analyze whether the structurally monotonous level will lead to meaningless grouping, i.e., similar structures can not reflect similar function. The analysis can help researchers realize how to select an appropriate structure fragment to proceed the local structure searching.

  On the other hand, we use several interesting cases to demonstrate the performance of fastCOPS. The fastCOPS can identify overlapping structural motifs and massive separate motifs through executing recursive truncation procedure. We also compare the capability of fastCOPS with other methods.

- Chapter 4 presents our conclusions and future work. In this study, we investigate the relationship of sequence-structure-function on protein fragments and purpose a novel framework–fastCOPS for local structure searching. The contribution of fastCOPS is can be used to identify structural motifs. Through expanding this application, researchers can further implement some tasks. For instance, protein structure prediction, fragment library design, antibody design, and drug design.

Chapter 2

Materials and Methods

In order to study the relationship of sequence-structure-function, we collect structural fragments those are functional annotated from PROSITE database. However, PROSITE is a sequence-based database. Our task is to validate whether these functional fragments are structure-conserved (they are sequence-conserved).

We use MAMMOTH to proceed the structural alignment on the fragments we selected. From the alignment results of intra-pattern (fragments belong to the same PROSITE pattern) and inter-pattern (fragments belong to distinct PROSITE pattern), we can observe the two distributions of structural alignment. Then, we design a re-clustering procedure to group structurally similar patterns. We intend to investigate the grouping circumstances and review the biological meaning PROSITE annotated to verify the relationship of structure-function.

Furthermore, we develop a quick and accurate tool for local conserved protein structure searching called "fastCOPS." The fastCOPS is a filter-and-refine framework. It integrates a quick screening method for pre-filtering and a accurate structural alignment method for refining. In addition, we embed the recursive truncation procedure on fastCOPS to enable the capability to identify multiple separate or overlapping fragments.

Finally, we select several interesting examples to demonstrate the performance of fastCOPS.

2.1 Structural Analysis of Patterns from PROSITE Database

PROSITE [13] is an annotated collection of motif descriptors dedicated to the identification of protein families and domains. These motifs usually embedded specific residues or regions that present important biological meaning and conserved in sequence and structure. We can describe that PROSITE is a sequence-based motif database, and maintains abundant and complete annotation of biological function. It robustly interpreted the relationship of sequence and function.

However, in order to study the relationship of sequence-structure-function, we are interested in investigating whether structural conservation reflect functional similarity. We collect structure fragments that corresponding with the patterns PROSITE defined to proceed structural analysis. Through observing the results of intra and inter-pattern structural alignment, we could evaluate whether structurally conserved property implies the function.

Furthermore, we design a re-clustering procedure based on structural similarity. By the results of re-clustered PROSITE pattern entries, we could review the documents PROSITE annotated to probe into the relationship of structure and function.

2.1.1 Collecting Patterns from PROSITE Database

We collect 1054 patterns from PROSITE database (Release 20.37) [14] for structural analysis. There are several entries of Protein Data Bank (PDB) that contain the structural fragments corresponding the sequence region described by each PROSITE pattern. Since the patterns PROSITE defined are sequence-based, not all patterns have corresponding 3D structure information. Each pattern we selected from PROSITE database has at least one 3D structure in current PDB.

From the 1054 patterns, we select structural fragments those sequence identity of each pattern are less than 90%. Owning to scientific purpose or technological limitations, researcher are allowed to store similar molecules those studied previously in PDB. However, we proceed the statistical analysis based upon non-redundant dataset will reduce the

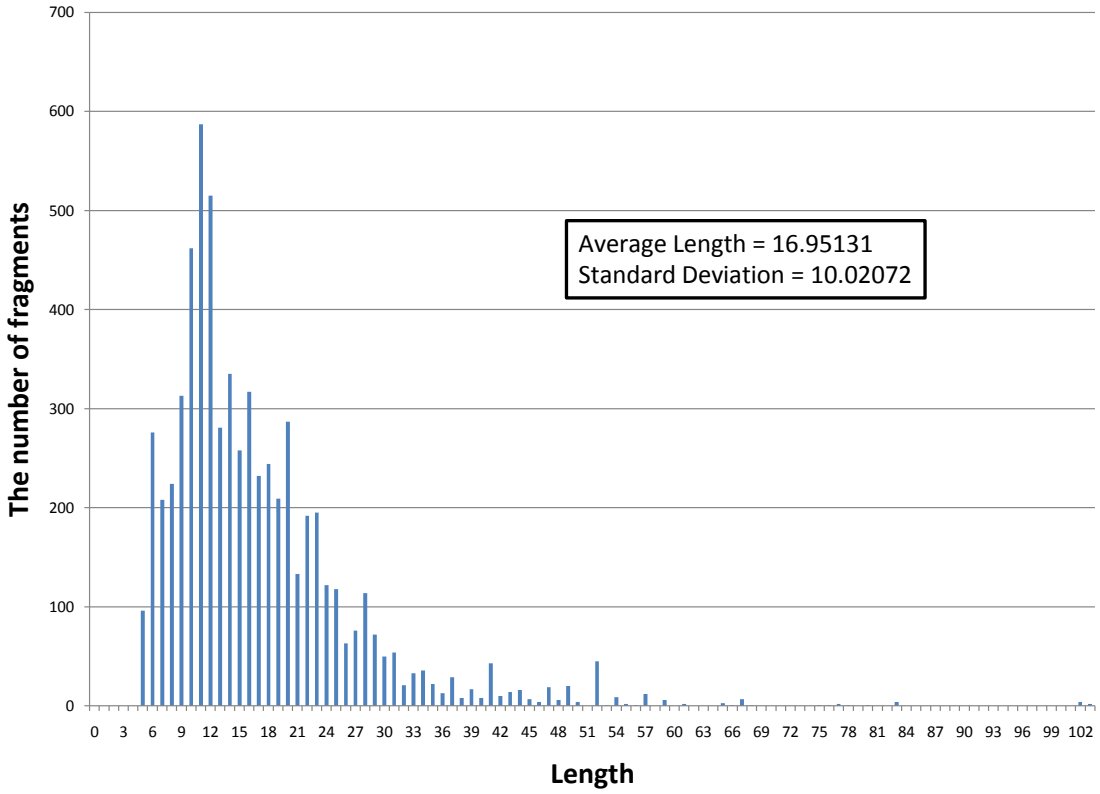

Figure 2.1: The mean length and standard deviation are 17 and 10, respectively. The range of pattern length comes within the scope of this study.

computational requirements and be more representative [15].

There are 6466 fragments we totally selected. As aforementioned, each fragment corresponds with the specific pattern description that PROSITE defined and eliminate redundancy. Figure 2.1 shows the length distribution of those fragments. The mean length and standard deviation are 17 and 10, respectively. The range of pattern length comes within the scope of this study.

2.1.2 Structural Alignment of PROSITE Patterns

There are total 1054 PROSITE patterns include total 6466 fragments we selected for proceeding structural alignment to observe the structurally conserved property. Each pattern has at least one structural fragments. We use these fragments to carry out pairwise alignment reciprocally. We select MAMMOTH [16] as structural alignment tool in this investigation.

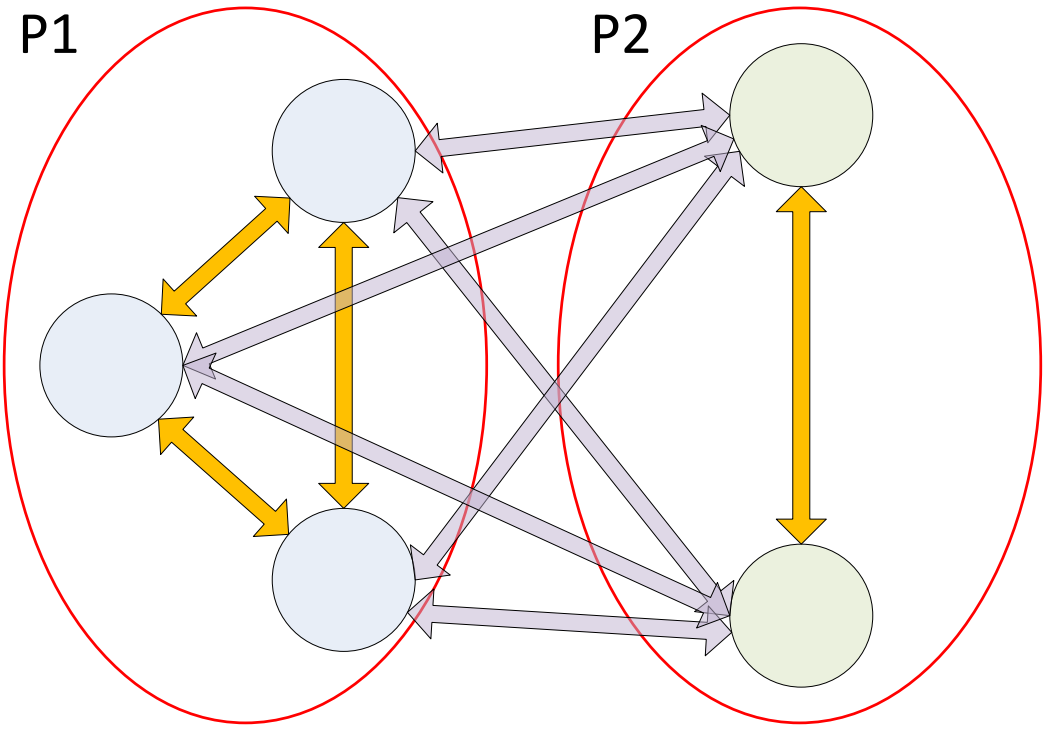

Figure 2.2: The sketch map of intra and inter pattern alignments.

When the pair of fragments belong to the same PROSITE pattern, the result of alignment will be aggregated as intra-pattern. On the other hand, when the pair of fragments belong to different patterns, the result of alignment will be aggregated as inter-pattern. Figure 2.2 shows the sketch map of intra and inter pattern alignments.

The number of reciprocal alignments of intra-pattern is the combination as $\binom{n}{2}$, where n is the number of fragments belong to the same PROSITE pattern. The total number of combinations of intra-pattern alignment is as follows:

$$\sum_{i=1}^{1054} \binom{n_i}{2} = 98226$$

We calculate the average RMSD (Root Mean Square Deviation) within the same pattern entries to present the structural conversed property of intra-pattern.

In addition, the number of reciprocal alignments of inter-pattern is the product of the of n and m, where n and m are the number of fragments of compared PROSITE pattern pair. The total number of combinations of inter-pattern alignment is as follows:

$$\sum_{i=1}^{1054} \frac{n_i(6466 - n_i)}{2} = 20803119$$

We also calculate the average RMSD within the combinations of distinct pattern entries to indicate the structural similarity of inter-pattern. The Figure 2.3 shows the RMSD probability distribution diagram of intra and inter-pattern. The diagram indicates that two distribution look like distinct. We use statistics method–Wilcoxon Rank-Sum Test to validate the fact.

Wilcoxon rank-sum test is a non-parametric test for assessing whether two independent samples of observations come from the same distribution. It is one of the best-known non-parametric significance tests. Through Wilcoxon rank-sum test, we reject the null hypothesis at the 5% significance level, i.e., the RMSD distributions of intra and inter-pattern are distinct distributions. Base on this fact, we would like to focus attention on exploring whether the two distributions imply the relationship of structure similarity and functional identity.

2.1.3 Structure-based Re-clustering Procedure

As Figure 2.3 indicates, some different PROSITE patterns have high structural similarity. In order to intend to investigate the relationship between protein structure and function, we design a clustering procedure to group these patterns with similar conformation. The structure-based clustering procedure will re-cluster the sequenced-based PROSITE pattern entries according to their mutual structural similarity. Figure 2.4 shows the flowchart of the re-clustering procedure.

The steps of the clustering procedure is as follows:

1. The inter-pattern pairwise alignment results are sorted by RMSD and put in queue.

2. If the queue is not empty, remove the head element of the queue and judge whether the results of the element crosses the threshold we defined. If the queue is empty, terminate the re-clustering procedure.

3. Assign patterns to the clusters

   (a) If both patterns of the element were got in step 2 have never grouped to any cluster, merge the two patterns to create a new cluster.

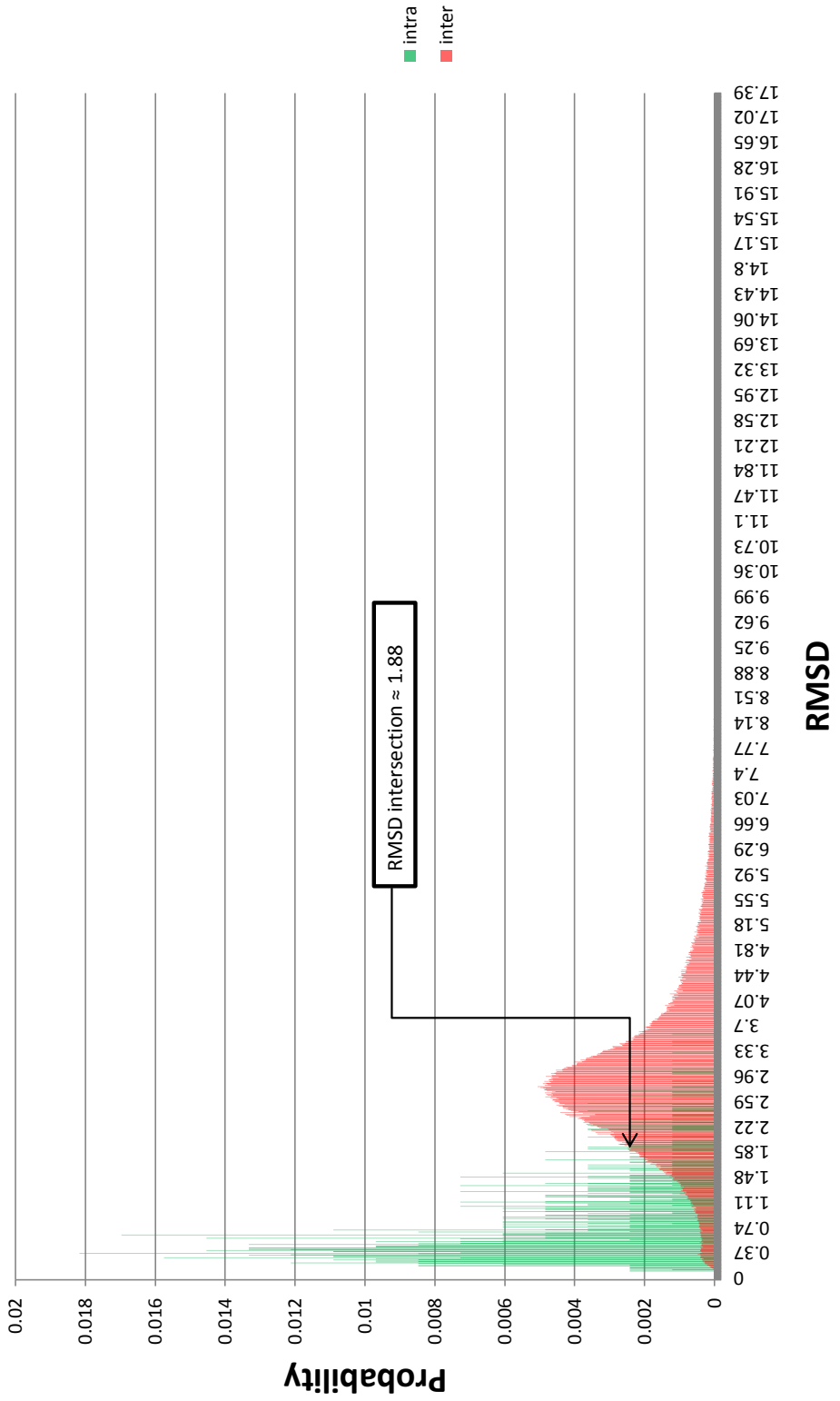

Figure 2.3: The RMSD probability distribution of intra-pattern and inter-pattern. At intra-pattern distribution, the mean, standard deviation, and the maximum are 0.98Å, 0.72Å, and 3.60Å, respectively. At inter-pattern distribution, the mean, standard deviation, and the maximum are 2.96Å, 1.21Å, and 17.39Å, respectively.

(b) If one pattern A of the element not existed in any cluster, one pattern B of the element already existed in a cluster, find all elements in queue those pattern pair include A and any pattern which belongs to the same cluster with B. When the results of all elements cross the threshold, add A to rebuild the existed cluster.

(c) Otherwise, two patterns of the element were already assigned to distinct clusters. It should find all elements in queue those pattern pair include any pattern which either belongs to the same cluster with one pattern of the element, or another. When the results of all elements cross the threshold, merge two existed clusters to rebuild a new cluster.

4. Go to step 2.

The threshold we defined as follows:

1. RMSD < 1.88

   Where the RMSD threshold we defined is according to the intersection of intra and inter-pattern distributions.

2. PSI (the percentage of structural identity) > 0.8, which is the percentage of residues aligned.

$$PSI = \frac{NALI}{Length}$$

   Where NALI is the Number of residue ALIgned, and Length is number of residues of the region PROSITE described. We use dual PSI to confirm that the percentage of aligned residues enough and the length of aligned two patterns are near enough. If dual PSI > 0.8, it implies the coverage > 0.8.

$$coverage = \frac{Length_A}{Length_B}$$

   Without loss of generality, where $Length_B$ is assumed equal or longer than $Length_A$.

Through the re-clustering procedure, several examples observed will suffice to show the structure–function relationship (discussed in Chapter 3). Hence, this observation motivate us to develop a structure search tool. We believe the ability to identify local conserved protein structure is important in predicting protein function.

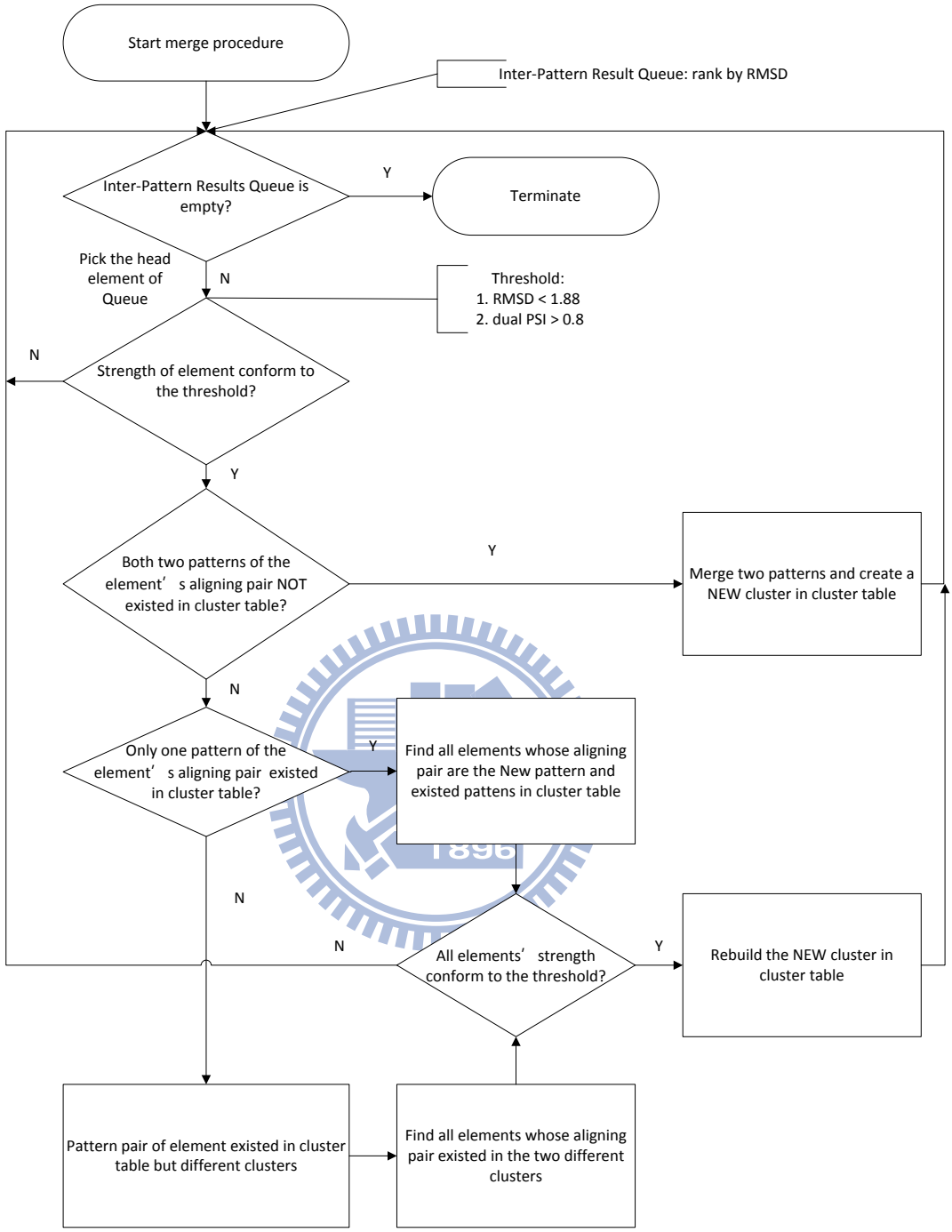

Figure 2.4: Flowchart of Re-Clustering procedure.

2.2 FastCOPS

We have developed a local protein structure search tool, fastCOPS (Fast Local Conserved Protein Structures Search Using 3D-BLAST [17], which has the features (e.g., robust statistical basis, effective search, and reliable database search capabilities) of BLAST to quickly screen structure databases for finding structural motifs. With the mechanism of recursive truncation, fastCOPS uses 3D-BLAST as an initial filter for screening databases and MAMMOTH [16] for detailed structure alignment.

2.2.1 FastCOPS Framework

The proposed framework integrate a quick screening method, 3D-BLAST, and an accurate structural alignment method, MAMMOTH, with an appropriate logic flow as shown in Figure 2.5 in a pseudo code style. In general, tools and methods currently available can be adopted in the fastCOPS framework as long as they are compatible with the design. The quick component should be able to accept a protein fragment as input, and the accurate component has to be capable of aligning partial structures with possible gap insertions. As aforementioned, our current implementation employs 3D-BLAST as the quick screening component and MAMMOTH as the accurate structural alignment component. The work flow is shown in Figure 2.6.

3D-BLAST

BLAST (Basic Local Alignment Search Tool) [3] is widely used tool for comparing primary biological sequence similarity. It's algorithm emphasizes speed to make searching task of huge genome databases efficient.

3D-BLAST is as fast as BLAST and calculates the statistical significance of an alignment to indicate the reliability of the prediction. In order to apply BLAST to structural alignment, 3D-BLAST used the structural alphabets that represent pattern profiles of the backbone fragments and then used them to represent protein structure databases as structural alphabet sequence databases (SADB). Structural alphabet substitution matrix (SASM) were developed on 3D-BLAST, which is used to replace default matrix for sequence alignment tool [17].

procedure FASTCOPS(Method$_{quick}$, Method$_{accurate}$, Query q)
  Result set $R \leftarrow$ execute Method$_{quick}(q)$
    for each item $r \in R$ **do**
      Structure alignment $r.align \leftarrow$ execute Method$_{accurate}(r)$
      if $r.align \neq \emptyset$ **then**
        Call RecursiveTruncation(Method$_{accurate}, r$)
      end if
    end for
    Report the structure alignment result on q
end procedure
procedure RecursiveTruncation(Method$_{accurate}$, Query q)
  $p \leftarrow q.truncation\_percentage$
  $q_{head} \leftarrow [q.start, q.align$ without the last $p\%]$ of q
  $q_{head}.align \leftarrow$ execute Method$_{accurate}(q_{head})$
  if $q_{head} \neq \emptyset$ **then**
    Call RecursiveTruncation(Method$_{accurate}, q_{head}$)
  end if
  $q_{tail} \leftarrow [q.align$ without the first $p\%, q.end]$ of q
  $q_{tail}.align \leftarrow$ execute Method$_{accurate}(q_{tail})$
  if $q_{tail} \neq \emptyset$ **then**
    Call RecursiveTruncation(Method$_{accurate}, q_{tail}$)
  end if
end procedure

Figure 2.5: The pseudo code of the fastCOPS for identifying structural motifs by searching protein structure databases.

Through the approach that translates protein 3D structure information to 1D sequence and the customized substitution matrix for structural alignment, 3D-BLAST provides efficiency on structure databases searching. In other words, the main advantage of 3D-BLAST is to reduce searching time of large structure database. It is a appropriate tool that can achieve filtering work.

MAMMOTH

MAMMOTH (Matching molecular models obtained from theory) [16] is a detailed structural comparison approach. It is sequence-independent, focuses on model C_α coordinates, and avoids reference to sequence or contact maps. The method is also capable of considering only portions of the protein.

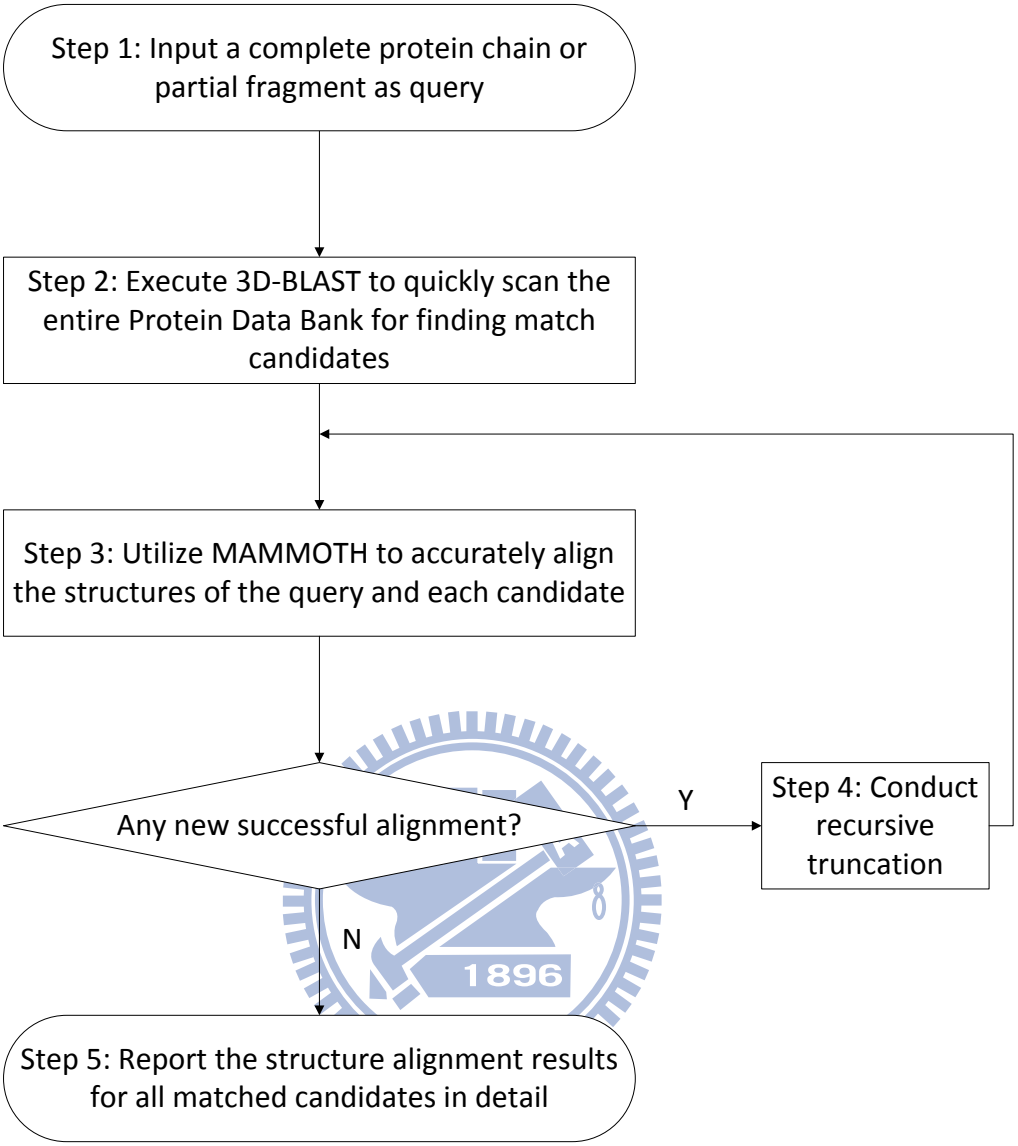

Figure 2.6: The workflow of the fastCOPS for identifying structural motifs by searching protein structure database.

Although MAMMOTH can provides detailed structural alignment, it is still a pairwise alignment tool. On huge databases searching task, using MAMMOTH is obviously impractical. In filter-and-refine approach, let MAMMOTH handle limited candidates from pre-filter method returned is a feasible scheme.

Recursive Truncation Procedure

For enhancing the sensitivity of local structure searching, we develop a recursive procedure to implement the discovery of multiple separate or overlapping fragment on a protein.

The concept of recursive truncation is giving the opportunity of alignment to other segments except the first matching region. When we use MAMMOTH to proceed pair-wise alignment of a complete chain and a short fragment, MAMMOTH only captures the optimal matching region once. However, in many cases, specific fragment may be multiple existence on a protein. To truncate the matching region and use the other segments to proceed aligning procedure is the thorough and robust approach to discover all possible fragments we interested. Figure 2.7 shows the sketch map of recursive truncation procedure.

Work Flow

At step 1. the user specifies the query by using a PDB code or a user-upload file in the PDB format, the chain, the range of residues, and the truncation percentage as the criteria. At step 2, fastCOPS executes the adopted quick method, i.e., 3D-BLAST, to quickly scan the entire PDB as a filter for comprehensively finding potential candidates which may match the query protein fragment. Next, MAMMOTH proceeds to conduct the detailed structural alignment between the query and each candidate returned by 3D-BLAST at

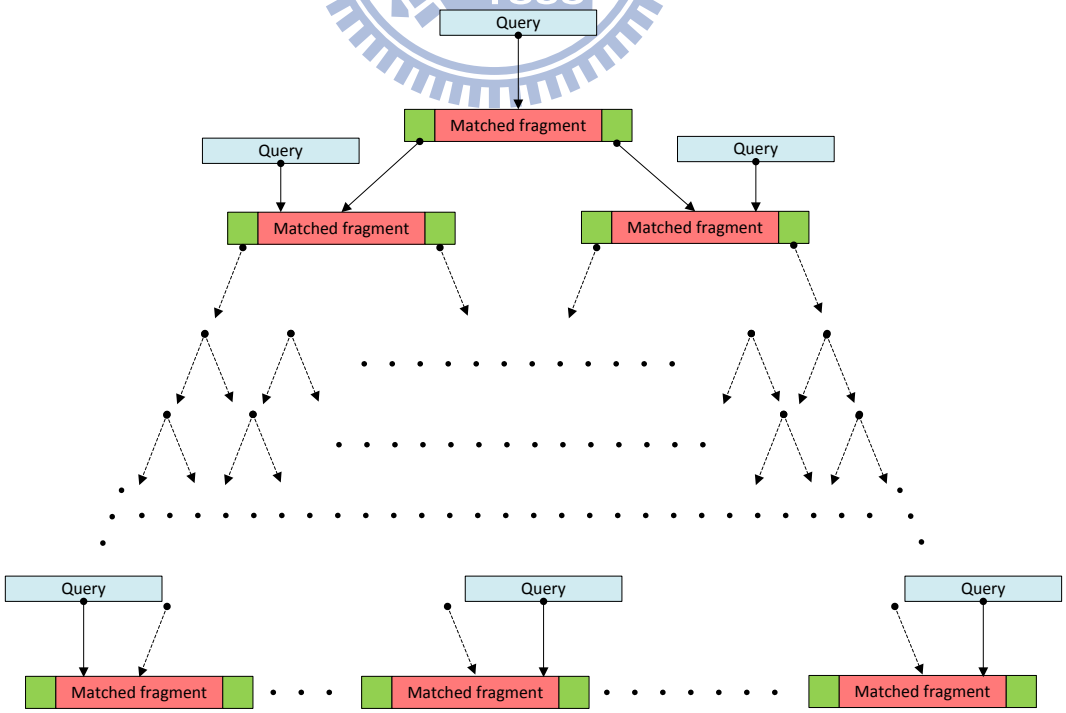

Figure 2.7: The sketch map of recursive truncation.

step 2 for accurate structural alignment. The recursive truncation procedure recursively refines the matching candidates reported by MAMMOTH until no more matches can be obtained. The framework design with recursive truncation enables fastCOPS to discover multiple separate or overlapping fragments that can be structurally aligned with the query.

In the recursive truncation procedure, we can set the truncation percentage to determine the level at which fastCOPS captures the multiple fragments that can match the query structure. If there is no need to conduct recursive queries, the zero value can be specified. If finding overlapping fragments is unnecessary, a vaule of 100 can be used. Otherwise, a value between 0 and 100 can be set according to the desirable query results.

Overall, fastCOPS performs two main steps to identify the protein structures similar to the query. 3D-BLAST was applied to quickly find the potential candidates as a filter, followed by the application of MAMMOTH to accurately align the structure. While aligning structures, MAMMOTH yields certain important statistics like PSI (the percentage of structural identity), NALI (number of residues aligned), Z-score, -ln(E) (alignment score, it is negative of the natural logarithm of the expected random value for that superimposition), and RMSD of the C_α atom position of the aligned residues between the query and the candidate. These statistics given by MAMMOTH quantify the quality of the resultant structural alignment.

2.2.2 Application

As significantly increasing in the number of protein crystal structures and the progress of structural genomics, identifying structural motifs from protein structures is one of the emergency tasks in structural bioinformatics.

Structural motif is fundamental for many applications, such as protein structure prediction, fragment library design, antibody design, and drug design. However, only a few of these methods have been designed for local structural motif search from large structure databases.

We can apply fastCOPS to achieve the task of structural motif search and identification. The fastCOPS has been evaluated on various structural motifs, including the treble clef finger motif, and the leucine-rich repeat motif.

In addition, we use a PROSITE pattern to demonstrate the capability that fastCOPS can find structurally conserved fragments but using sequence alignment tool will hardly identify them.

Treble clef finger motif

We first illustrate the search results of fastCOPS with the query of the treble clef finger motif [18]. This structural motif consists of a zinc knuckle followed by a loop, a β-hairpin, and an α-helix, and is characterized by the distinct structural arrangement of these elements. This case is used to demonstrate the capability of finding overlapping fragments of fastCOPS.

Leucine-rich repeat motif

For the second structural motif, the leucine-rich repeat (LRR) motif was used as the query structure to demonstrate that the fastCOPS is capable of identifying massive structurally similar fragments on a protein. LRR occurs in proteins ranging from viruses to eukaryotes. Most LRRs are 20-29 amino acids long and present in a number of proteins with diverse functions. The primary function of these structural motifs [19] appears to provide a versatile structural framework for the formation of protein–protein interactions [20].

PROSITE pattern PS00853: PLP attachment site

We use PROSITE pattern PS00853–PLP attachment site to scan entire PDB. PLP attachment site has a conserved residue, lysine side chain. It reacts with PLP to form Schiff base [21].

Chapter 3

Results and Discussions

3.1 Statistical Analysis of PROSITE Pattern Comparisons

We select 1054 PROSITE patterns (include total 6466 structure fragments) to proceed structural pairwise alignment mutually. When the aligned pair of fragments belong to the same PROSITE pattern, the result will be aggregated as intra-pattern, otherwise, they will be aggregated as inter-pattern. Figure 2.3 shows the RMSD probability distribution of intra-pattern and inter-pattern. We locally enlarge the distribution diagram to identify the intersection conveniently and show as broken line graph on Figure 3.1.

From the distribution diagram, we can observe the situation of intra-pattern present smaller RMSD (the mean, standard deviation, and the maximum are 0.98Å, 0.72Å. 3.60Å, respectively). The result of intra-pattern reflects the structure conserved property (they are sequence conserved). On the other hand, the inter-pattern present weaker RMSD distribution (the mean, standard deviation, and the maximum are 2.96Å, 1.21Å, and 17.39Å), naturally. The intersection of two distributions is at 1.88Å.

However, the question which we must consider is why the pairwise alignment results of distinct patterns display structure similarity. Then, they present the structure conserved property whether they also reflect functional similarity. Full discussion will be presented in the next section.

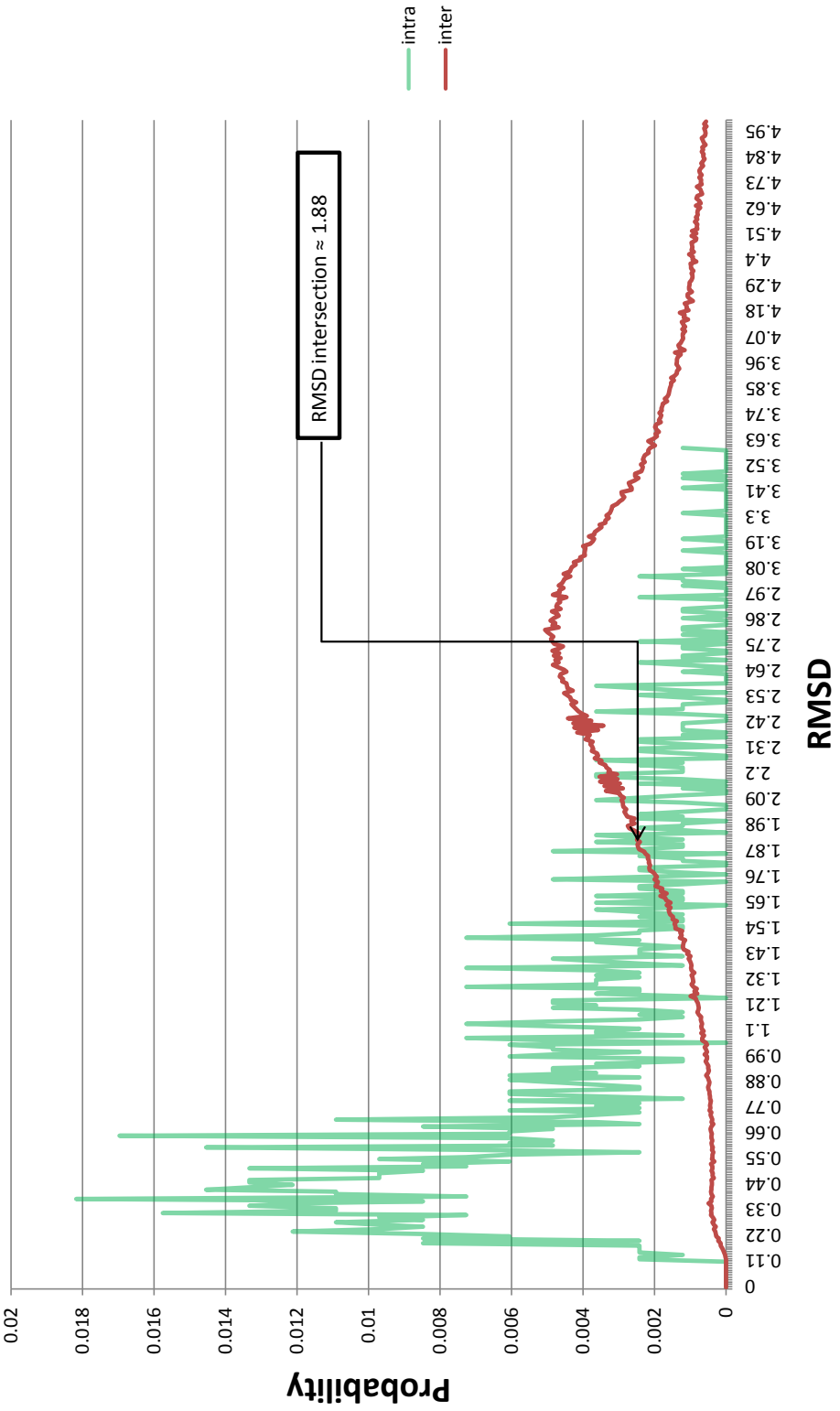

Figure 3.1: The RMSD probability distribution of intra-pattern and inter-pattern (locally enlarged). The intersection of two distributions is at 1.88.

3.2 Interpretation of Re-clustering Results

We re-cluster the PROSITE patterns according to their structural similarity. Using the intersection (RMSD = 1.88Å) of intra and inter RMSD distributions to be the threshold, the procedure (mentioned in Chapter 2) grouped 410 PROSITE patterns (among total 1054 patterns) to 147 clusters.

From the results of re-clustering, we discover several cases that present significant structure similarity and also reflect the functional similarities by reviewing the PDOC PROSITE provided. The cases validate the relationship of structure and function, i.e., structural conservation implies functional identity.

We list three cases to be illustrations in the following paragraph.

Protein Kinases Signatures

PROSITE pattern entries: PS00108, PS00109, and PS01245 are grouped by the re-clustering procedure. PS00108 and PS00109 are described by the same PROSITE documentation entry (PDOC00100: Protein kinases signatures and profile). In addition, PS01245 is described in PDOC00958, that described the pattern entry is highly conserved central part of protein family RIO1, ZK632.3 and MJ0444.

The description of PDOC00100 is as follows:

" Eukaryotic protein kinases [22, 23, 24, 25, 26] are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases."

Their regular expression are as follows:

1. PS00108: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3)
2. PS00109: [LIVMFYC]-A-[HY]-x-D-[LIVMFY]-[RSTAC]-D-PF-N-[LIVMFYC](3)
3. PS01245: [LIVMY]-[VI]-H-[GA]-D-[LF]-[SN]-E-[FY]-N-x-[LIVM]

We also review the literature of PDB code 1ZP9, 1ZTF, and 1ZTH [27] that correspond with PROSITE pattern PS01245 described. The region of the PROSITE pattern PS01245

matched is the catalytic loop on the protein. Especially to deserve to be mentioned, the family of 1ZP9, 1ZTF, and 1ZTH is the atypical protein kinases (aPKs). The atypical kinases is not significantly identified with eukaryotic protein kinases (ePKs) in sequence but contain kinase signature.

Figure 3.2 shows the conformations of protein structural fragments corresponding with pattern PS00108, PS00109, PS01245 and the multiple structural alignment.

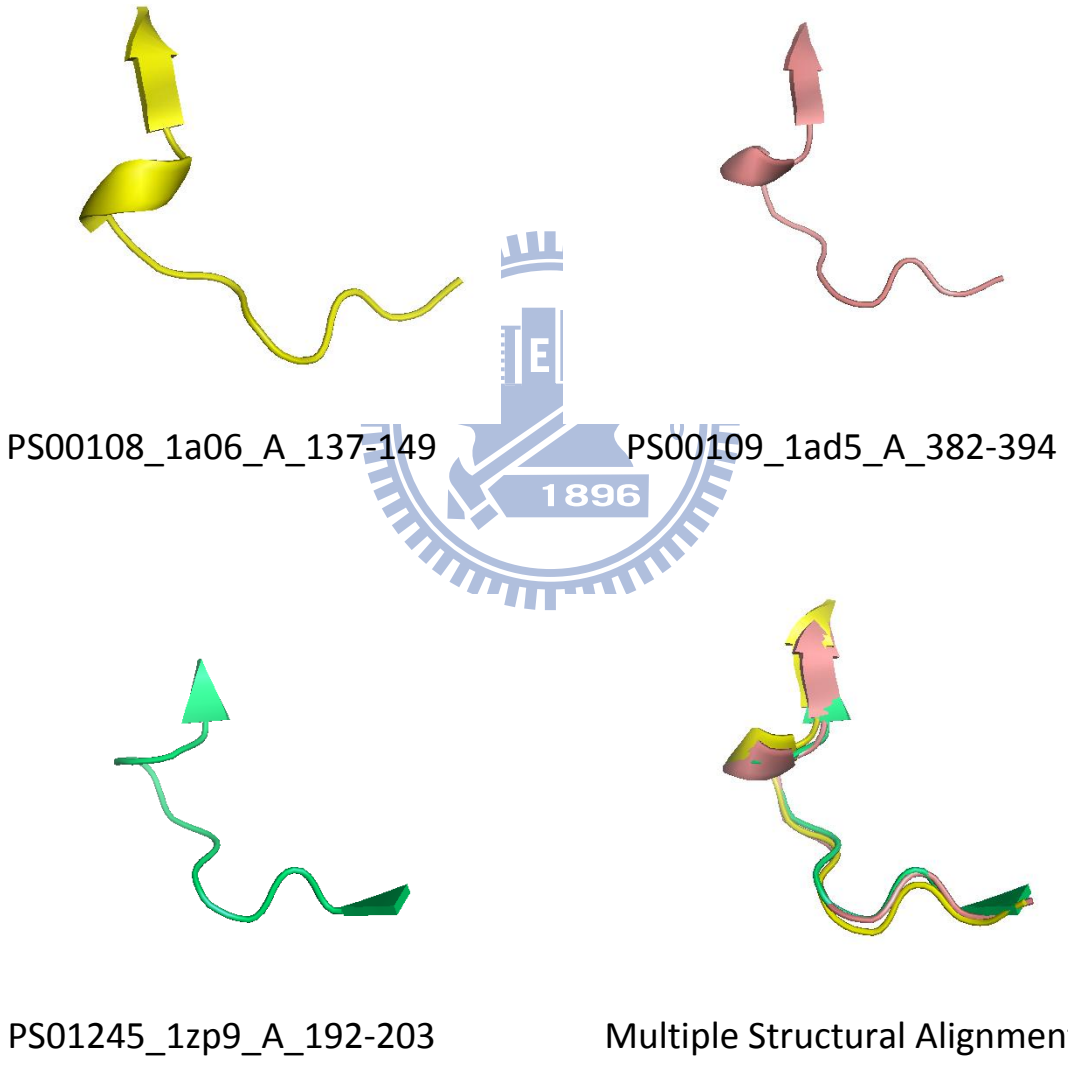

Figure 3.2: Conformations of Protein Kinases Signatures on 1A06-A, 1AD5-A, and 1ZP9-A. We selected one protein fragment from each PROSITE pattern entry and shows the multiple structural alignment.

Serine Proteases Signatures

PROSITE pattern entries: PS00135 and PS00673 are grouped by the re-clustering procedure. PS00135 and PS00673 are described by PDOC00124 (Serine proteases, trypsin family, signatures and profile) and PDOC00571 (Serine proteases, V8 family, active sites), respectively.

The description of PDOC00124 is as follows:

"The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine."

And, the description of PDOC00571 is as follows:

"A number of prokaryotic proteases have been shown [28, 29] to be evolutionary related; ***their catalytic activity is provided by a charge relay system similar to that of the trypsin family of serine proteases*** *but which probably evolved by independent convergent evolution."*

According to the two PROSITE documentation, we can realize the two patterns are functional similarity. Their regular expression are as follows:

1. PS00135: [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH]

2. PS00673: T-x(2)-[GC]-[NQ]-S-G-S-x-[LIVM]-[FY]

They are obviously distinct in sequence, and may hardly be identified by sequence alignment tool mutually. However, in structural aspect, they are similar and present functional similarity.

Figure 3.3 shows the conformations of protein structural fragments corresponding with pattern PS00135, PS00673 and the multiple structural alignment.

Pyridoxal-phosphate (PLP) attachment site

PROSITE pattern entries: PS00853 and PS00096 are grouped by the re-clustering procedure. PS00853 and PS00096 are described by PDOC00667 (Beta-eliminating lyases

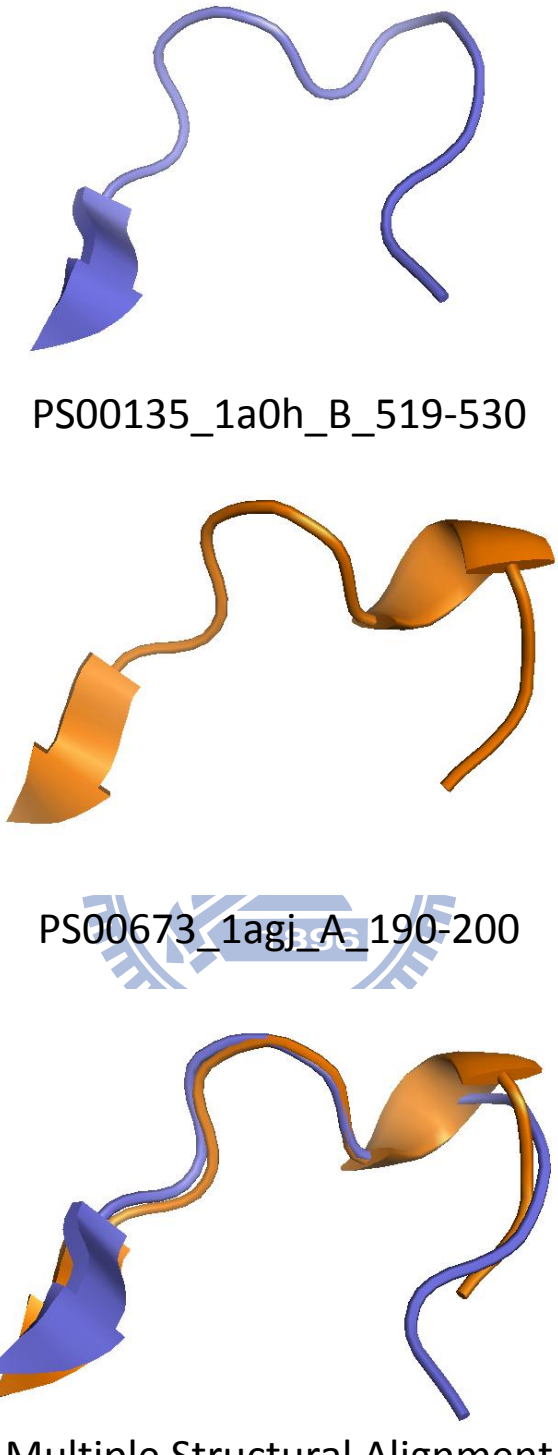

Figure 3.3: Conformations of Serine Proteases Signatures on 1A0H-B, and 1AGJ-A. We selected one protein fragment from each PROSITE pattern entry and shows the multiple structural alignment.

pyridoxal-phosphate attachment site) and PDOC00090 (Serine hydroxymethyltransferase pyridoxal-phosphate attachment site), respectively.

PS00853 and PS00096 are related pyridoxal-phosphate dependent homotetrameric enzymes and pyridoxal-phosphate containing enzyme, respectively. Both of them have attachment site whose central section of the sequence is lysine residue is used to attach pyridoxal-phosphate group.

The regular expression are as follows:

1. PS00853: [YV]-x-D-x(3)-M-S-[GA]-K-**K**-D-x-[LIVMF]-[LIVMAG]-x-[LIVM]-G-G
2. PS00096: [DEQHY]-[LIVMFYA]-x-[GSTMVA]-[GSTAV]-[ST]-[STVM]-[HQ]-**K**-[STG]-[LFMI]-x-[GAS]-[PGAC]-[RQ]-[GSARH]-[GA]

Except the central lysine (K) residue, the other part of sequence identity seems not significant.

Figure 3.4 shows the conformations of protein structural fragments corresponding with pattern PS00096, PS00853 and the multiple structural alignment.

Merged Monotonous Structures

After surveying results the re-clustering procedure executed, we discover some cases of merged PROSITE pattern entries can not present the functional identity or similarity due to monotonous structures.

For instance, there is a strange case that combined 17 PROSITE pattern entries. As Table 3.1 shows, they do not present the functional identity even if they have similar structures. According to the encoding rule of 3D-BLAST, on the encoded structural fragments, we evaluate the continuous three structural alphabets represent helix (A, Y, B, C, and D) and the structural alphabets represent strand (E, F, and H) as α-helix and β-strand, respectively [8]. The second column in the table means the percentage of α-helix or β-strand on the protein fragments (average value of the same pattern entry). We can observe that the most of merged pattern entries in this case have at least 70% helix or strand. We selected one fragment from each pattern entry to present the merged conformations as Figure 3.5 shows. They are all monotonous α-helix. Because of the regularity

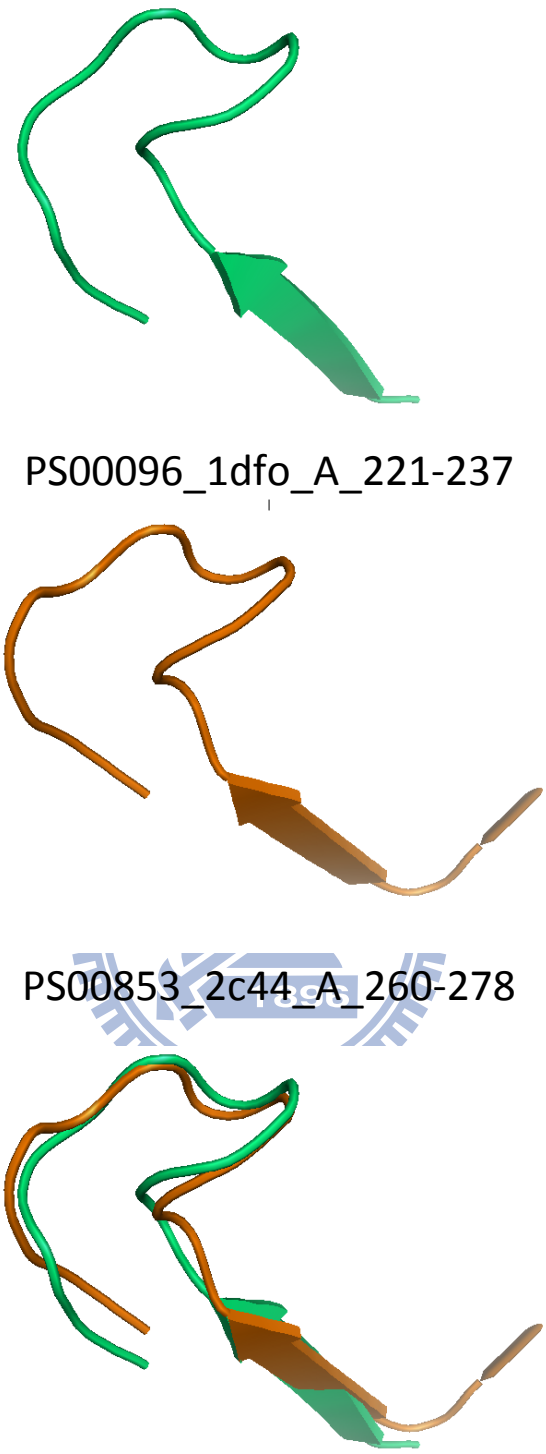

Figure 3.4: Conformations of Pyridoxal-phosphate (PLP) attachment site on 1DFO-A, and 2C44-A. We selected one protein fragment from each PROSITE pattern entry and shows the multiple structural alignment.

#PS	%Helix/Strand	Functional description
PS00654	0.8461538461538461	PRD domain signature
PS01039	0.7678571428571429	Bacterial extracellular solute-binding proteins, family 3 signature
PS00819	0.6666666666666666	Dps protein family signature 2
PS00211	0.7733333333333332	ABC transporters family signature
PS01001	0.8571428571428571	Succinate dehydrogenase cytochrome b subunit signature 2
PS00468	0.8809523809523809	Eukaryotic cobalamin-binding proteins signature
PS00444	0.9120879120879122	Polyprenyl synthetases signature 2
PS01297	0.7666666666666666	FLAP/GST2/LTC4S family signature
PS00946	1.0	Cathelicidins signature 1
PS00715	1.0	Sigma-70 factors family signature 1
PS00624	0.5333333333333333	GMC oxidoreductases signature 2
PS00648	0.7333333333333333	Bacterial ribonuclease P protein component signature
PS01111	0.7555555555555555	RNA polymerases K / 14 to 18 Kd subunits signature
PS00153	0.7799145299145299	ATP synthase gamma subunit signature
PS00954	0.8571428571428572	Imidazoleglycerol-phosphate dehydratase signature 1
PS01327	0.7142857142857143	Large-conductance mechanosensitive channels mscL family signature
PS00950	0.7692307692307693	Bacterial rhodopsins signature 1

Table 3.1: A strange case, 17 PROSITE patterns are merged by structural similarity, but they do not present functional identity or similarity.

of helix, they have excellent value of RMSD while proceeding structural alignment. We consider this kind of structural similarity is meaningless, and could not reflect functional identity.

Furthermore, we review all PROSITE PDOC of pattern entries after executing re-clustering procedure. We verify the functional identity of each pattern entry. If a pattern entry is merged with others, we would check whether their function described by PROSITE possess identity or similarity. Figure 3.6 shows the relationship of the helix/strand percentage of the protein fragment and functional identity.

We can observe that the higher helix/strand percentage accompanies the lower functional identity. When the helix/strand percentage is higher than 70%,and the functional identity is even lower than 30%. This result means when the protein fragment has high

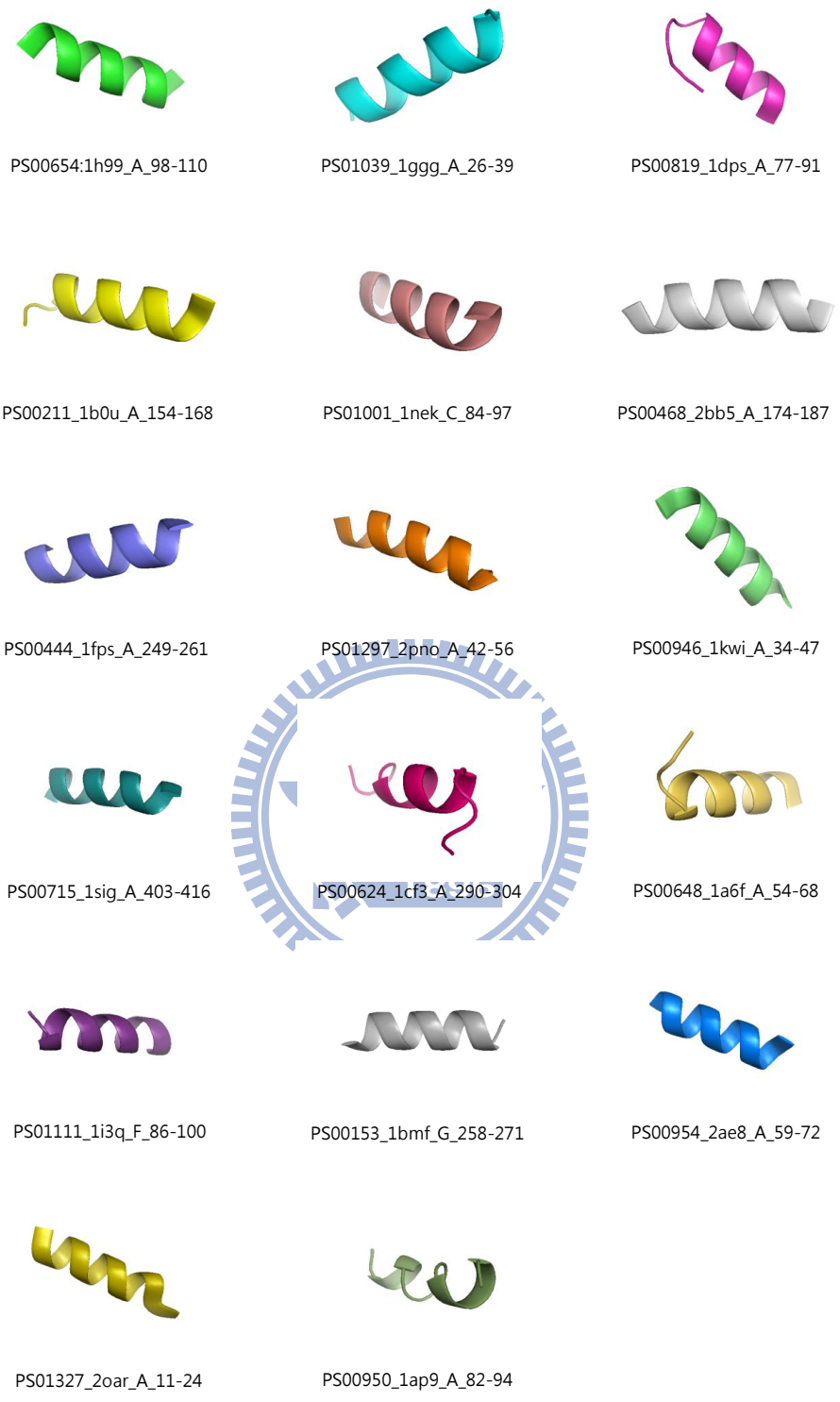

Figure 3.5: A strange case, 17 PROSITE patterns are merged by structural similarity, but they do not present functional identity or similarity.

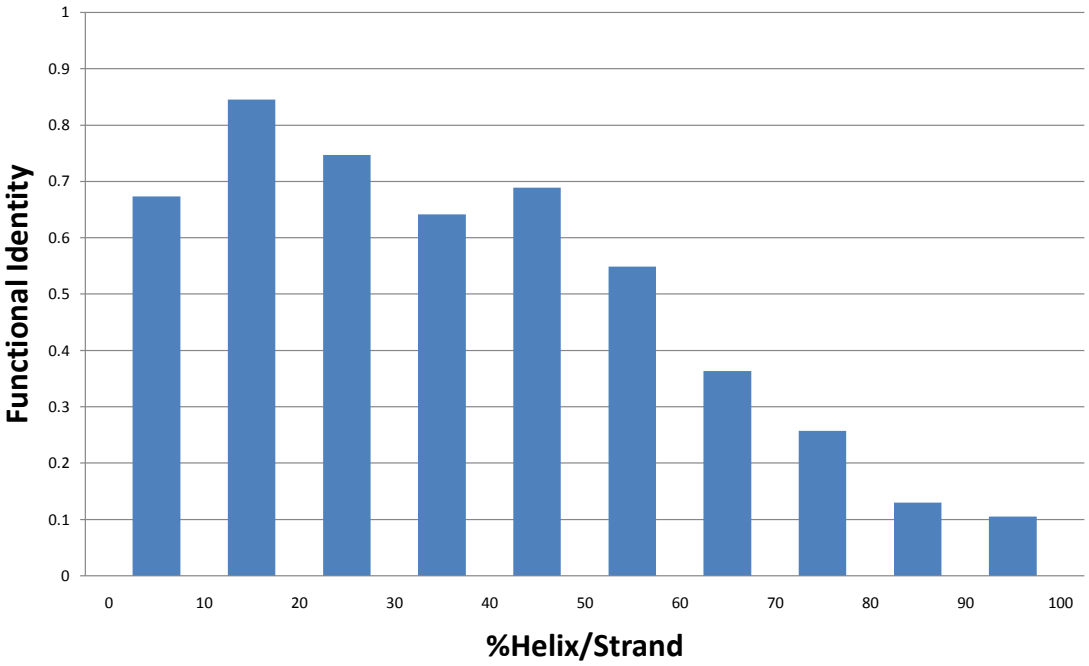

Figure 3.6: The relationship of functional identity and percentage of helix or strand, the higher helix/strand percentage accompanies the lower functional identity.

proportion of regular structure (helix/strand), it hardly presents structurally distinctive feature, then it will easily be confused with other structures, which also present regular and monotonous conformation.

On the other hand, when the proportion of regular structure in protein fragment is lower than 50%, the functional identity is around 70%. This observation can provide suggestion that how to select an appropriate protein fragment as query when we apply structural fragment searching tool like fastCOPS to scan databases.

3.3 FastCOPS Searching Cases

The fastCOPS is developed for finding local conserved structure (structural motif) on proteins. We select two examples to demonstrate the performance of overlapping motifs and multiple separate motifs, respectively. In addition, we try to use a PROSITE pattern entry–PS00853 (Pyridoxal-phosphate (PLP) attachment site) to scan the entire PDB.

Through fastCOPS, we successfully found several cases whose sequence similarity are low, and even could not be identify by sequence alignment.

3.3.1 Treble Clef Finger Motif

The first case is used to demonstrate the capability of finding overlapping motifs. We use the zinc-binding segment (H190-D235) of phosphatidylinositol-3-phosphate binding FYVE domain of vps27p protein from yeast (PDB code 1VFY-A) as the query [30], the fastCOPS found 25 similar local structures, which are overlapping zinc finger motifs in the FYVE domain of 12 proteins (Table 3.2), by searching on PDB. This query domain contains two overlapping zinc finger motifs (Figure 3.7(A)).

For instance, the fastCOPS identifies the two overlapping zinc finger motifs of Hrs protein (PDB code 1DVP-A) [31], where one segment is H178-R219 and the other is R162-Q195. The overlapping part is between residues H178-Q195. Among these 25 similar segments, the sequence identities of 14 segments, which are often unable to be identified by sequence alignment tools, are less than 25%. The distribution between the sequence identity and RMSD values is shown in Figure 3.8. Figure 3.7(B) shows the multiple structural alignment of the query segment and structure motifs of the second zinc finger of four proteins, including Hrs (PDB code 1DVP-A, orange), endosomal antoantigen 1 (PDB code 1JOC-B, blue), and two structural genomics targets (PDB code 2YW8-A, purple and 1X4U-A, cyan).

PDB code	Fragment	SCOP domain family	RMSD (Å)	Seq. Identity
1vfy	A(190-234)	vps27p protein	0	100%
1dvp	A(178-219)	FYVE domain (Hrs)	0.59	50%
1joc	B(1372-1410)	FYVE domain (Eea1)	1.04	49%
1joc	A(1372-1410)	FYVE domain (Eea1)	1.1	49%
2yw8	A(36-74)	FYVE domain [a]	1.16	51%
2yqm	A(43-83)	FYVE domain [a]	1.32	49%
1z2q	A(38-82)	FYVE domain [a]	1.42	42%
1wfk	A(26-67)	FYVE domain	1.45	26%
1hyj	A(26-64)	FYVE domain (Eea1)	1.63	49%
1hyi	A(26-64)	FYVE domain (Eea1)	1.86	49%
1x4u	A(31-79)	FYVE domain [a]	2.11	33%
1wim	A(6-43)	UbcM4-interacting protein 4	3.02	17%
1zbd	B(109-149)	FYVE domain	3.14	22%
1z2q	A(22-56)	FYVE domain [a]	3.23	21%
2yw8	A(20-53)	FYVE domain [a]	3.26	21%
1hyj	A(10-43)	FYVE domain (Eea1)	3.35	18%
1joc	B(1356-1389)	FYVE domain (Eea1)	3.39	18%
2yqm	A(27-60)	FYVE domain [a]	3.39	24%
1x4u	A(15-48)	FYVE domain [a]	3.4	15%
1hyi	A(10-43)	FYVE domain (Eea1)	3.45	18%
1joc	A(1356-1389)	FYVE domain	3.49	18%
1dvp	A(162-195)	FYVE domain (Hrs)	3.54	15%
1vfy	A(174-207)	vps27p protein	3.57	12%
1wfk	A(10-43)	FYVE domain	3.67	15%
1zbd	B(92-126)	FYVE domain	3.68	21%

[a] The domain is a structural genomics target and its SCOP domain is obtained by using the fastSCOP [32].

Table 3.2: The fastCOPS search results using phosphatidylinositol-3-phosphate binding FYVE domain of vps27p protein from yeast (PDB code 1VFY-A (H190-D235)) as the query. There are 25 similar structures belonging to 13 distinct protein chains found by fastCOPS.

3.3.2 Leucine-Rich Repeat Motif

The second case is used to demonstrate the capability of finding multiple separate motifs. We use the protein fragment, T54-N77 in PDB code 1XEC-A, as the LRR query [33]. The fastCOPS found 504 similar local structures which belong to the top 50 distinct protein chains (filtered by 3D-BLAST) and match the query. The sequence identities of 229 segments among total results are less than 25%.

The fastCOPS successfully identified 23 canonical LRRs as well as one irregular LRR (N-terminal cap region) in 1ZIW-A (Figure 3.9 and Table 3.3). The sequence identities and RMSD values of these 23 identified LRRs range 13%–39% and 0.67Å–2.71Å, respectively. The distribution between the sequence identities and RMSD values of using the LRR motif as query is shown in Figure 3.10. 1ZIW is called human Toll-like receptor 3 (TLR3) [34]. We can obseve that the aligned length of the found irregular LRR is much shorter than that of the canonical LRRs except for LRR12 and LRR20 as shown in Table 3.3. Because 17 out of the 23 human TLR3 LRRs have the canonical 24-residue motif, and only LRR12 and LRR20 have insertions longer than 5 residues, the aligned lengths for LRR12 and LRR20 are quite short compared to that of other canonical LRRs. Such a situation indicates that fastCOPS can provide accurate structure alignments by adopting MAMMOTH.

3.3.3 PROSITE pattern: PS00853–PLP attachment site

In this case, we selected the fragment that PROSITE pattern PS00853–Pyridoxal-phosphate attachment site described to demonstrate the capability of finding structural conserved PLP attachment site. We use PDB code 1C7G, chain A, and the range of residues from 247 to 265 as query. Table 3.4 shows the part results that RMSD < 1.88Å. The fastCOPS successfully identified several cases that not match PROSITE regular expression, i.e., if researchers use sequence alignment tool will hardly identify them. Figure 3.11 shows the distribution of RMSD and sequence identity.

PDB code	Fragment	Structural alignment	RMSD (Å)	Seq. Identity
1c7g	A(247-264)	YADGCTMSGKKDCLVNIG	0	100%
1c7g	B(247-264)	YADGCTMSGKKDCLVNIG	0.08	100%
1c7g	C(247-264)	YADGCTMSGKKDCLVNIG	0.09	100%
1c7g	D(247-264)	YADGCTMSGKKDCLVNIG	0.1	100%
2vlf	B(247-264)	YADGCTMSGKKDCLVNIG	0.12	100%
2vlf	A(247-264)	YADGCTMSGKKDCLVNIG	0.16	100%
2vlh	A(247-264)	YADGCTMSGKKDCLVNIG	0.16	100%
2ez2	A(247-264)	YADGCTMSGKKDCLVNIG	0.19	100%
2ez2	B(247-264)	YADGCTMSGKKDCLVNIG	0.19	100%
1tpl	A(247-264)	YADGCTMSGKKDCLVNIG	0.28	100%
1tpl	B(247-264)	YADGCTMSGKKDCLVNIG	0.3	100%
2c44	B(260-277)	YADMLAMSAKKDAMVPMG	0.49	56%
2c44	A(260-277)	YADMLAMSAKKDAMVPMG	0.5	56%
2c44	C(260-277)	YADMLAMSAKKDAMVPMG	0.5	56%
2c44	D(260-277)	YADMLAMSAKKDAMVPMG	0.5	56%
2oqx	A(260-277)	YADMLAMSAKKDAMVPMG	0.55	56%
2v0y	A(260-277)	YADMLAMSAKKDAMVPMG	0.55	56%
2v1p	A(260-277)	YADMLAMSAKKDAMVPMG	0.55	56%
1bjo[a]	A(188-205)	RYGVIYAGAQKNIGPAGL	1.43	6%
1ax4[a]	B(256-274)	YADALTM.SAKDDPLLNIGG	1.6	50%
1ax4[a]	C(256-274)	YADALTM.SAKDDPLLNIGG	1.61	50%
1ax4[a]	D(256-274)	YADALTM.SAKDDPLLNIGG	1.61	50%
1ax4[a]	A(256-274)	YADALTM.SAKDDPLLNIGG	1.62	50%
2z9u[a]	B(187-204)	KADIYVTGPNKCLGAPPG	1.62	22%
2e7i[a]	B(199-216)	GADFIVGSGHKSMAASGP	1.64	28%
1lw4[a]	C(189-206)	YADSVMFCLSKGLCAPVG	1.67	28%
1lw5[a]	C(189-206)	YADSVMFCLSKGLCAPVG	1.68	28%
2ez1[a]	A(247-265)	YADGCT.MSGKDDCLVNIGG	1.72	50%
2ez1[a]	B(247-265)	YADGCT.MSGKDDCLVNIGG	1.72	50%
2tpl[a]	B(247-265)	YADGCT.MSGKDDCLVNIGG	1.72	50%
2jis[a]	B(295-312)	RADSVAWNPHKLLAAGLQ	1.74	17%
2jis[a]	A(295-312)	RADSVAWNPHKLLAAGLQ	1.75	17%
2tpl[a]	A(247-265)	YADGCT.MSGKDDCLVNIGG	1.75	50%
2vlh[a]	B(247-265)	YADGCT.MSGKDDCLVNIGG	1.75	50%

Table 3.4: The fastCOPS search results using PROSITE pattern–PS00853 (Pyridoxal-phosphate (PLP) attachment site) as query. [a] The attachment site could not be identified by PROSITE regular expression.

3.4 Comparison with Other Methods

A few methods were specifically developed to perform local structure search task. We select FF [10] and PAST [11] to proceed the comparisons with the fastCOPS by cases, because their configurations of input and output resemble the fastCOPS.

FF (Fragment Finder) is a web-based interface. It has 25% and 90% non-homologous protein chains databases as searching space. Users can input a specific PDB-ID, chain, and the interested region of residues to proceed the search task of structural motif. FF has several parameters can be tuned, which include searching database, X-ray diffraction /NMR model, tolerance level of the conformation angles, ... and so on.

PAST (Polypeptide Angle Suffix Tree) is also a web service. It uses the entire PDB as searching space, but dose not use a filtered subset as representatives. The input configuration of PAST like FF. PAST also has parameters can be tuned, which include torsion angles type, tolerance level of the conformation angles, and C_α RMSD cutoff.

We use the two cases (TCF and LRR motif) aforementioned as inputs to perform the comparisons. The tolerance level of the conformation angles is the main factor influencing the searching results on FF and PAST. The default tolerance level of FF is 5°. However, we could not get any results by using the default parameters on FF with the two cases. It seems too severe. The default tolerance level on PAST is ± 3 coding interval (70°). So, we also set the tolerance level to $(\pm)35^\circ$ on FF for the fair comparing conditions.

The parameters we set on FF as follows:

- Sequence: Search for structurally similar fragments having any sequence (default)
- Structure solved by: X-ray Diffraction (default)
- Structures based on: 90% Non homologous identity
- Tolerance: 35°

The parameters we set on PAST as follows:

- Tolerance: ± 3 (70°; default)

- RMSD cutoff: 2.5 (Å; default)

In the first case, we use the TCF motif (PDB code 1VFY-A: 190-235) as input on FF and PAST. FF could not present any result and PAST identified the only one fragment itself. On the other hand, our approach–fastCOPS, can identify 25 similar structures belonging to 13 distinct protein chains.

In the second case, we use the LRR motif (PDB code 1XEC-A: 54-77) as input on FF and PAST. FF identified 2 similar structures belonging to 1 protein chain. And, PAST identified 77 similar structures belonging to 22 distinct protein chains. On the other hand, fastCOPS can identify 504 similar structures belonging to 50 distinct protein chains. There are 363 similar structures those RMSD $\leq$ 2.5Å belonging to 30 distinct protein chains among the total results of fastCOPS.

Especially to deserve to be mentioned, FF, PAST, and fastCOPS all identified the protein chain: 1OZN-A. FF, PAST, and fastCOPS identified 2, 3, 8 fragments on 1OZN-A those RMSD $\leq$ 2.5Å, respectively. Table 3.5 shows the identified structural fragments by FF, PAST, and fastCOPS, respectively.

Through the comparison with FF and PAST, it reveals that the capability of local structure search on fastCOPS is comparable and competitive within the same kind of approaches.

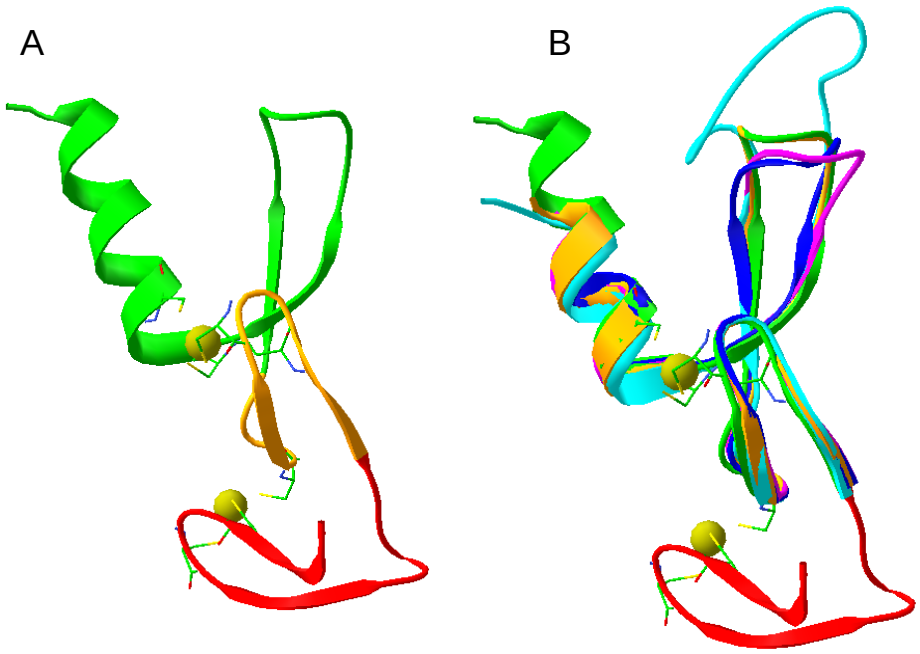

Figure 3.7: (A) Two overlapping zinc finger motifs of the query structure FYVE domain (1VFY-A). The first zinc finger (S173-C207) is colored in red and the second one (H190-D235) is colored in green. The overlapping segment is colored in orange. (B) A multiple structural alignment of the query segment and hit segments of the second zinc finger in four proteins, including Hrs (1DVP-A, orange), endosomal autoantigen 1 (1JOC-B, blue), two structural genomics targets (2YW8-A, purple; 1X4U-A, cyan).

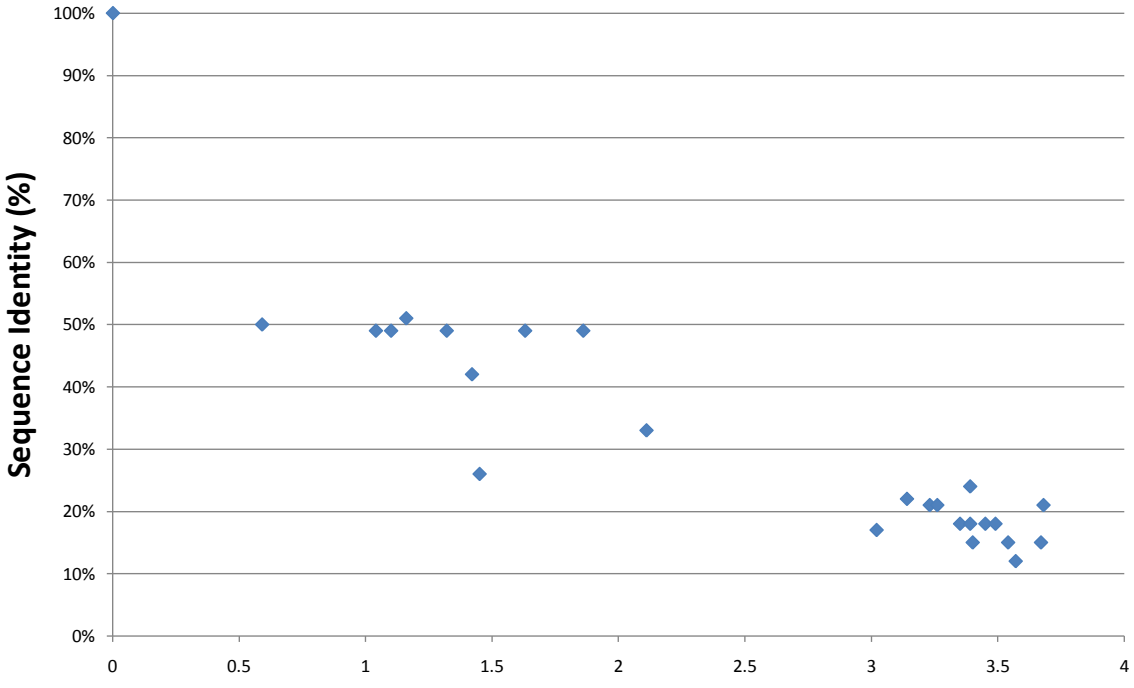

Figure 3.8: The distribution between the sequence identities and RMSD values of 25 similar structures using phosphatidylinositol-3-phosphate binding FYVE domain of vps27p protein from yeast (PDB code 1VFY-A (H190-D235)) as the query.

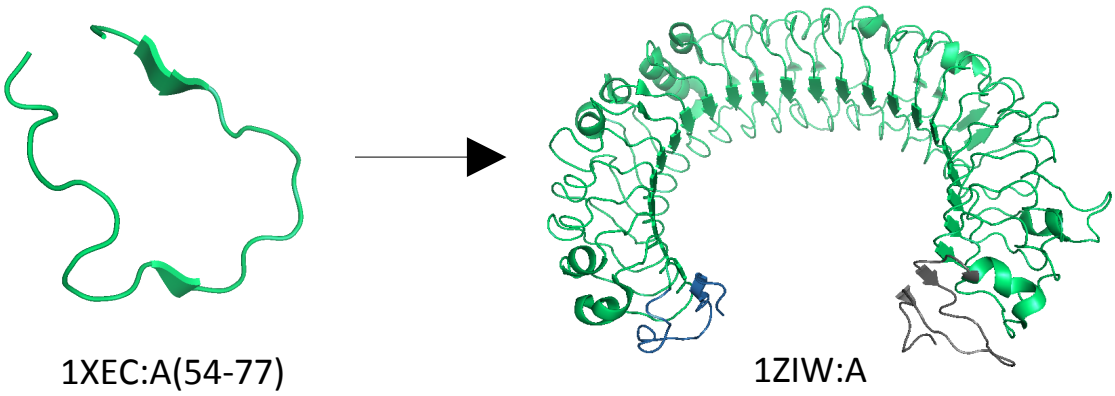

Figure 3.9: 1ZIW-A can be found by the fastCOPS with query 1XEC-A (T54-N77).

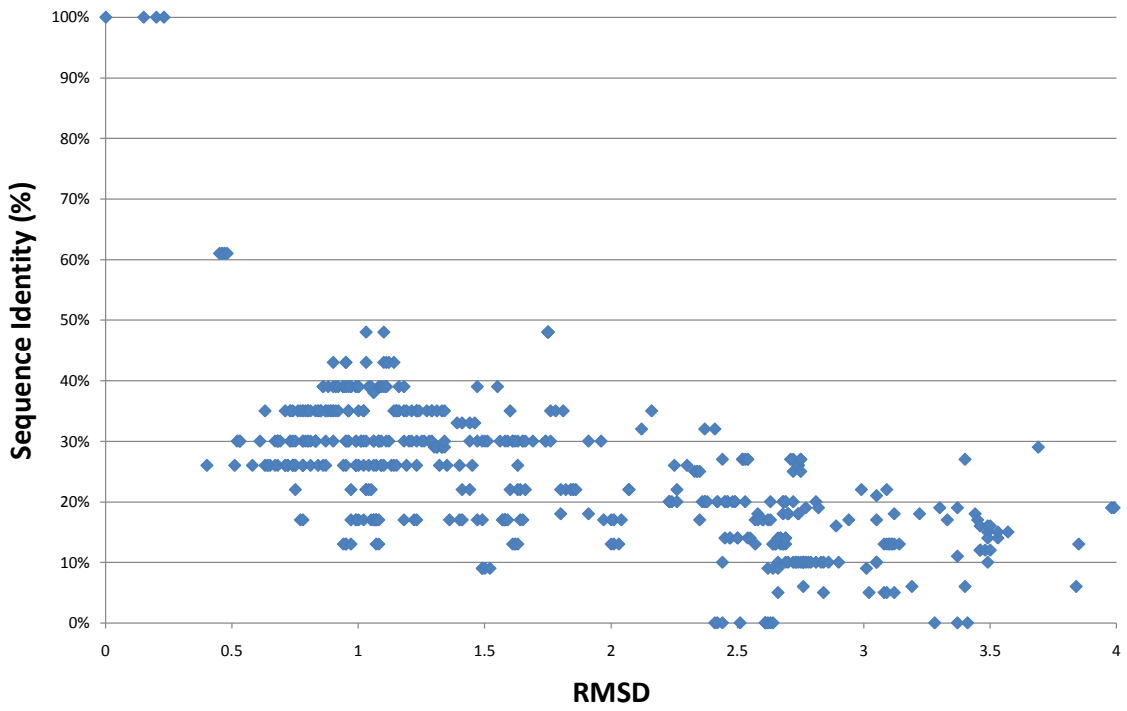

Figure 3.10: The distribution between the sequence identity and RMSD of 504 similar structures using the leucine-rich repeat (LRR) motif (T54-N77 in PDB code 1XEC-A) was used as the query structure.

Residue No. in 1ZIW-A	Structural Alignment	Aligned Length	RMSD	Description
32-51	VSHEVADCSHLKLTQVPDDLPT...	20	2.54	N-terminal cap region
53-75	PTNITVLNLTHNQLRRLPAANFTRYSQLT	23	1.07	Canonical LRR 1
77-99	QLTSLDVGFNTISKLEPELCQKLP	23	1.34	Canonical LRR 2
101-123	LPMLKVLNLQHNELSQLSDKTFAFCTNLT	23	0.94	Canonical LRR 3
125-147	NLTELHLMSNSIQKIKNNPFVKQKNLI	23	0.98	Canonical LRR 4
149-171	NLITLDLSHNGLSSTKLGTQVQLE	22	2.31	Canonical LRR 5
173-195	LENLQELLLSNNKIQALKSEELDIFANSS	23	1.40	Canonical LRR 6
199-221	NSSLKKLELSSNQIKEFSPGCFHAIGRLF	23	1.56	Canonical LRR 7
223-246	RLFGLFLNNVQLGPSLTEKLCLELANT	23	3.21	Canonical LRR 8
250-272	NTSIRNLSLSNSQLSTTSNTTFLGLKWT	23	1.48	Canonical LRR 9
276-298	WTNLTMLDLSYNNLNVVGNDSFAWLP	23	0.94	Canonical LRR 10
300-322	LPQLEYFFLEYNNIQHLFSHSLHGLFNVR	23	0.86	Canonical LRR 11
324-355	NVRYLNLKRSFTKQLPKIDDFSFQWLK	21	2.73	Canonical LRR 12
357-379	LKCLEHLNMEDNDIPGIKSNMFTGLINLK	23	0.67	Canonical LRR 13
381-405	NLKYLSLSNSFTSLRTLTNETFVSLAHS	23	2.71	Canonical LRR 14
409-431	HSPLHILNLTKNKISKIESDAFSWLGHLE	23	0.81	Canonical LRR 15
433-456	HLEVLDLGLNEIGQELTGQEWRGLENIF	23	1.33	Canonical LRR 16
458-480	NIFEIYLSYNKYLQLTRNSFALVPSLQ	23	1.33	Canonical LRR 17
482-506	SLQRLMLRRVALKNVDSSPSPFQPLR	23	2.51	Canonical LRR 18
508-530	LRNLTILDLSNNNIANINDDMLEGLEKLE	23	1.24	Canonical LRR 19
532-562	KLEILDLQHNNLARLWKHANPGGPIYFLKGLS	17	1.61	Canonical LRR 20
564-586	LSHLHILNLESNGFDEIPVEVFKDLFELK	23	0.86	Canonical LRR 21
588-610	ELKIIDLGLNNLNTLPASVFNNQVSLK	23	1.02	Canonical LRR 22
612-634	SLKSLNLQKNLITSVEKKVFGPAFRNL	23	1.87	Canonical LRR 23

Table 3.3: The 23 canonical LRRs and one irregular LRR on PDB Code 1ZIW-A are identified by fastCOPS with query 1XEC-A(T54-N77: TALLDLQNNKITEIKDGDFKNLKN).

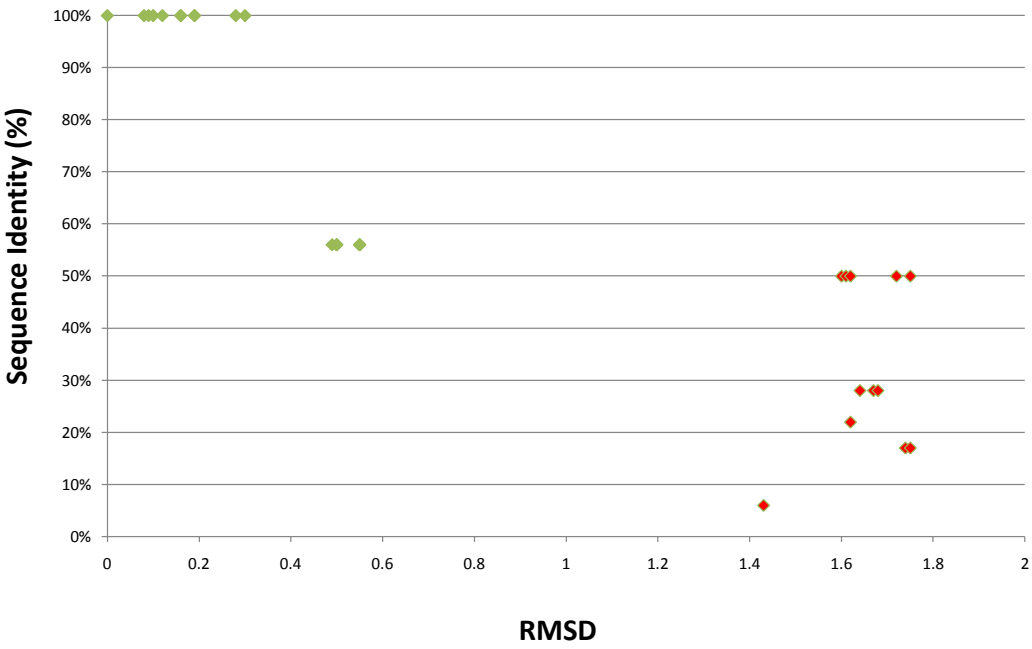

Figure 3.11: The distribution between the sequence identity and RMSD of 34 similar structures using PLP attachment site (PDB code 1C7G-A: 247-265) was used as the query structure. The green points means the found fragments match PROSITE regular expression and the red points means the found fragments not match PROSITE regular expression.

PDB code	Fragment	FF	PAST	fastCOPS
1ozn	A(59-81)			✓
1ozn	A(83-105)	✓	✓	✓
1ozn	B(107-130)			✓
1ozn	A(132-154)			✓
1ozn	A(156-178)			✓
1ozn	A(180-202)	✓	✓	✓
1ozn	A(204-226)		✓	✓
1ozn	A(228-250)			✓

Table 3.5: The identified similar structures on 1OZN-A, using 1XEC-A: 54-77 as input.

Chapter 4

Conclusions

4.1 Summary

In order to investigate the relationship of sequence-structure-function, we analyze the structurally conserved properties of PROSITE pattern. PROSITE is an annotated collection of motifs, which usually embedded specific residues or regions to perform biological function. These patterns are conserved in sequence and structure.

From observing the distribution of structural similarity (represented by RMSD) of intra and inter-pattern alignment, we observed that some distinct PROSITE patterns have high structural similarity.

Then, we design a re-clustering procedure to group PROSITE patterns with similar conformations. We discover several cases to validate the fundamental principle in protein science, i.e., structure leads to function. However, the results of re-clustering by structural similarity also occur some strange cases due to structurally monotonous. We reviewed the structural property (percentage of regular secondary structural element) of all grouped clusters and whether them reflect functional identity or similarity. When the proportion of regular structure in protein fragment is higher than 70%, and the functional identity is even lower than 30%. This result give us the hint that how to select an appropriate structure fragment to proceed local structure search for avoiding false positive.

In addition, we develop a novel framework–fastCOPS composed of 3D-BLAST, for quick screening, MAMMOTH, for detailed structural alignment, and recursive truncation, for refining search results. By recursive truncation procedure, we can find overlapping or multiple separate structural motifs.

With the fastCOPS, researchers can rapidly scan the entire protein structure databases to find the proteins containing local conserved structures similar to the given structure. Through demonstrating two examples, we perform the capability of finding overlapping and multiple separate structural motifs. Finally, we select a PROSITE pattern to scan entire PDB. The result present that fastCOPS can find structurally conserved and functionally similar fragments, but using sequence alignment tool seems hardly be achieved.

4.2 Major Contributions and Future Work

Major Contributions

Through surveying the PROSITE database, we validate the relationship of sequence-structure-function and discover the issue when proceeding short fragment search may cause high false positive due to monotonous structure. We reviewed all grouped PROSITE patterns with structural similarity and the corresponding PROSITE documentations to conclude the relationship between monotonous level of structure and the functional identity that grouped patterns present. The result may give us the hint that how we select a structure fragment as query to scan databases when using local structure searching tool.

On the other hand, we develop a robust and solid local conserved structure searching tool–fastCOPS. The fastCOPS can deliver structure alignment results accurately, quickly, and thoroughly. In fact, fastCOPS also provides high flexibility for researchers. According to the researcher's requirement or favorite, they can select the region of protein with any length, even a complete chain to proceed structural database searching and structure alignment. The framework of fastCOPS is not limited to only perform local structure searching and identification.

We believe that the fastCOPS can effectively identify structural motifs and can be a useful service for annotating the functions of novel structures.

Future Work

As a future work, we can further improve the ability of fastCOPS by changing the two main components of filter-and-refine framework when more novel methods will be published. Besides, the current fastCOPS uses single structure fragment as query to find similar

structures. Through appropriately expending the framework, fastCOPS may be applied to use multiple structure fragments as queries to find separate, discontinuous structural motifs or spatially conserved environments.

Bibliography

[1] G. A. Petsko and D. Ringe, *Protein Structure and Function*. New Science Press, 2004.

[2] P. E. Bourne and H. Weissig, *Structural Bioinformatics*. Wiley-Liss, 2003.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[4] C. Chothia and A. M.Lesk, "The relation between the divergence of sequence and structure in proteins," *The EMBO Journal*, vol. 5, no. 4, pp. 823–826, 1986.

[5] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path," *Protein Engineering*, vol. 11, no. 9, pp. 739–747, 1998.

[6] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of Molecular Biology*, vol. 233, pp. 123–138, September 1993.

[7] T. Madej, J.-F. Gibrat, and S. H. Bryant, "Threading a database of protein cores," *PROTEINS: Structure, Function, and Genetics*, vol. 23, pp. 356–369, 1995.

[8] C.-H. Tung, J.-W. Huang, and J.-M. Yang, "Kappa-alpha plot derived structural alphabet and blosum-like substitution matrix for rapid search of protein structure database," *Genome Biology*, vol. 8, p. R31, March 2007.

[9] S.-Y. Ku and Y.-J. Hu, "Protein structure search and local structure characterization," *BMC Bioinformatics*, vol. 9, no. 349, 2008.

[10] P. Ananthalakshmi, C. K. Kumar, M. Jeyasimhan, K. Sumathi, and K. Sekar, "Fragment finder: a web-based software to identify similar three-dimensional structural motif," *Nucleic Acids Research*, vol. 33, pp. W85–W88, 2005.

[11] H. Taubig, A. Buchner, and J. Griebsch, "Past: fast structure-based searching in the pdb," *Nucleic Acids Research*, vol. 34, pp. W20–W23, 2006.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[13] C. J. A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher, "Prosite: A documented database using patterns and profiles as motif descriptors," *Briefing in Bioinformatics*, vol. 3, no. 3, pp. 265–274, September 2002.

[14] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B. A. Cuche, E. de Castro, C. Lachaize, P. S. Langendijk-Genevaux, and C. J. A. Sigrist, "The 20 years of prosite," *Nucleic Acids Research*, pp. 1–5, November 2007.

[15] M. Moorhouse and P. Barry, *Bioinformatics, biocomputing and Perl: an introduction to bioinformatics computing skills and practice*. Wiley, 2004.

[16] A. R. Ortiz, C. E. Strauss, and O. Olmea, "Mammoth (matching molecular models obtained from theory): An automated method for model comparison," *Protein Science*, pp. 2606–2621, November 2002.

[17] J.-M. Yang and C.-H. Tung, "Protein structure database search and evolutionary classification," *Nucleic Acids Research*, vol. 34, no. 13, pp. 3646–3659, 2006.

[18] N. V. Grishin, "Treble clef finger–a functionally diverse zinc-binding structural motif," *Nucleic Acids Research*, vol. 29, no. 8, pp. 1703–1714, 2001.

[19] B. Kobe and J. Deisenhofer, "A structural basis of the interactions between leucine-rich repeats and protein ligands," *Nature*, vol. 374, pp. 183–186, 1995.

[20] B. Kobe and A. V. Kajava, "The leucine-rich repeat as a protein recognition motif," *Current Opinion in Structural Biology*, vol. 11, pp. 725–732, 2001.

[21] D. Whitford, *Proteins–Structure and Function*. Willey, 2005.

[22] S. K. Hanks and T. Hunter, "Protein kinases 6. the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification," *The FASEB Journal*, vol. 9, pp. 576–596, May 1995.

[23] T. Hunter, "Protein kinase classification," *Methods in enzymology*, vol. 200, pp. 3–37, 1991.

[24] S. K. Hanks and A. M. Quinn, "Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members," *Methods in enzymology*, vol. 200, pp. 38–62, 1991.

[25] S. K. Hanks, "Eukaryotic protein kinases," *Current Opinion in Structural Biology*, vol. 1, pp. 369–383, 1991.

[26] S. K. Hanks, A. Quinn, and T. Hunter, "The protein kinase family: conserved features and deduced phylogeny of the catalytic domains," *Science*, vol. 241, pp. 42–52, 1988.

[27] N. LaRonde-LeBlanc, T. Guszczynski, T. Copeland, and A. Wlodawer, "Structure and activity of the atypical serine kinase rio1," *the FEBS Journal*, vol. 272, pp. 3698–3713, 2005.

[28] S. J. Dancer, R. Garratt, J. Saldanha, H. Jhoti, and R. Evans, "The epidermolytic toxins are serine proteases," *FEBS Letters*, vol. 268, no. 1, pp. 129–132, July 1990.

[29] C. J. Bailey and T. P. Smith, "The reactive serine residue of epidermolytic toxin a," *Biochemical Journal*, vol. 269, pp. 535–537, 1990.

[30] S. Misra and J. H. Hurley, "Crystal structure of a phosphatidylinositol 3-phosphate-specific membrane-targeting motif, the fyve domain of vps27p," *Cell*, vol. 97, pp. 657–666, May 1999.

[31] Y. Mao, A. Nickitenko, X. Duan, T. E. Lloyd, M. N. Wu, H. Bellen, and F. A. Quiocho, "Crystal structure of the vhs and fyve tandem domains of hrs, a protein involved in membrane trafficking and signal transduction," *Cell*, vol. 100, pp. 447–456, February 2000.

[32] C.-H. Tung and J.-M. Yang, "fastscop: a fast web server for recognizing protein structural domains and scop superfamilies," *Nucleic Acids Research*, pp. 1–6, May 2007.

[33] P. G. Scott, P. A. McEwan, C. M. Dodd, E. M. Bergmann, P. N. Bishop, and J. Bella, "Crystal structure of the dimeric protein core of decorin, the archetypal small leucine-rich repeat proteoglycan," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 44, pp. 15 633–15 638, November 2004.

[34] J. Choe, M. S. Kelker, and I. A. Wilson, "Crystal structure of human toll-like receptor 3 (tlr3) ectodomain," *Science*, vol. 309, pp. 581–585, June 2005.

Appendices

Re-Clustering PROSITE Patterns by Structural Similarity

#C	#PS	%Helix/Strand	Description
1	PS00959	1.0	Histone H3 signature 2
1	PS00326	1.0	Tropomyosins signature
1	PS00304	0.7	Small, acid-soluble spore proteins, alpha/beta type, signature 1
1	PS00226	0.8472222222222222	Intermediate filaments signature
1	PS00019	0.8500000000000001	Actinin-type actin-binding domain signature 1
1	PS00345	0.9047619047619049	Ets-domain signature 1
1	PS00517	0.7777777777777778	Ribonuclease III family signature
1	PS00449	1.0	ATP synthase a subunit signature
1	PS00142	0.8566666666666661	Neutral zinc metallopeptidases, zinc-binding region signature
1	PS01041	0.6	Stathmin family signature 2
1	PS00865	0.38425925925925924	Ubiquitin-activating enzyme active site
1	PS00767	0.2777777777777778	Tetrahydrofolate dehydrogenase/cyclohydrolase signature 2
2	PS01050	0.7272727272727273	YjeF C-terminal domain signature 2
2	PS00312	1.0	Glycophorin A signature
2	PS00774	0.9090909090909091	Chitinases family 19 signature 2
2	PS00119	1.0	Phospholipase A2 aspartic acid active site
2	PS00796	0.8977272727272728	14-3-3 proteins signature 1
2	PS00613	0.9090909090909091	Osteonectin domain signature 2
2	PS00137	0.5619834710743802	Serine proteases, subtilase family, histidine active site
2	PS00138	0.4772727272727271	Serine proteases, subtilase family, serine active site
2	PS00664	0.30303030303030304	Vinculin repeated domain signature
2	PS00435	0.49090909090909085	Peroxidases proximal heme-ligand signature
2	PS00337	0.30303030303030304	Beta-lactamase class-D active site
3	PS00654	0.8461538461538461	PRD domain signature

Continued...

#C	#PS	%Helix/Strand	Description
3	PS01039	0.7678571428571429	Bacterial extracellular solute-binding proteins, family 3 signature
3	PS00819	0.6666666666666666	Dps protein family signature 2
3	PS00211	0.7733333333333332	ABC transporters family signature
3	PS01001	0.8571428571428571	Succinate dehydrogenase cytochrome b subunit signature 2
3	PS00468	0.8809523809523809	Eukaryotic cobalamin-binding proteins signature
3	PS00444	0.9120879120879122	Polyprenyl synthetases signature 2
3	PS01297	0.7666666666666666	FLAP/GST2/LTC4S family signature
3	PS00946	1.0	Cathelicidins signature 1
3	PS00715	1.0	Sigma-70 factors family signature 1
3	PS00624	0.5333333333333333	GMC oxidoreductases signature 2
3	PS00648	0.7333333333333333	Bacterial ribonuclease P protein component signature
3	PS01111	0.7555555555555555	RNA polymerases K / 14 to 18 Kd subunits signature
3	PS00153	0.7799145299145299	ATP synthase gamma subunit signature
3	PS00954	0.8571428571428572	Imidazoleglycerol-phosphate dehydratase signature 1
3	PS01327	0.7142857142857143	Large-conductance mechanosensitive channels mscL family signature
3	PS00950	0.7692307692307693	Bacterial rhodopsins signature 1
4	PS00029	1.0	Leucine zipper pattern
4	PS00812	0.763157894736842	Glycosyl hydrolases family 8 signature
4	PS00663	0.7619047619047619	Vinculin family talin-binding region signature
4	PS00706	0.9157894736842106	Prion protein signature 2
4	PS01260	0.6190476190476191	Apoptosis regulator, Bcl-2 family BH4 motif signature
4	PS00957	0.5681818181818181	NAD-dependent glycerol-3-phosphate dehydrogenase signature
4	PS00520	0.5714285714285714	Interleukin-10 family signature
5	PS00436	0.8833333333333334	Peroxidases active site signature
5	PS00513	0.6499999999999999	Adenylosuccinate synthetase active site
5	PS00068	0.7472527472527473	Malate dehydrogenase active site signature
5	PS00208	0.8333333333333334	Plant hemoglobins signature
5	PS00500	0.75	Thymosin beta-4 family signature

Continued...

#C	#PS	%Helix/Strand	Description
5	PS00933	0.5384615384615384	FGGY family of carbohydrate kinases signature 1
5	PS00353	0.8055555555555557	HMG box A DNA-binding domain signature
5	PS00498	0.6166666666666667	Tyrosinase and hemocyanins CuB-binding region signature
5	PS00523	0.6201923076923077	Sulfatases signature 1
5	PS01291	0.6923076923076923	NAD:arginine ADP-ribosyltransferases signature
6	PS00547	0.7222222222222222	Transglutaminases active site
6	PS01206	0.6190476190476191	Amiloride-sensitive sodium channels signature
6	PS00880	0.513157894736842	Acyl-CoA-binding (ACB) domain signature
6	PS00590	0.6107843137254902	LIF / OSM family signature
6	PS00470	0.5199404761904762	Isocitrate and isopropylmalate dehydrogenases signature
7	PS00368	0.7882352941176471	Ribonucleotide reductase small subunit signature
7	PS00540	0.7894736842105264	Ferritin iron-binding regions signature 1
7	PS00417	0.96	Synaptobrevin signature
7	PS00797	0.9224400871459696	14-3-3 proteins signature 2
7	PS00237	0.8544117647058822	G-protein coupled receptors family 1 signature
7	PS00360	0.7368421052631579	Ribosomal protein S9 signature
7	PS00549	0.8823529411764706	Bacterioferritin signature
8	PS00593	0.6363636363636364	Heme oxygenase signature
8	PS00327	0.5416666666666667	Bacterial rhodopsins retinal binding site
8	PS00531	0.5	Ribonuclease T2 family histidine active site 2
8	PS00367	0.15	Biopterin-dependent aromatic amino acid hydroxylases signature
9	PS00036	0.9851190476190477	Basic-leucine zipper (bZIP) domain signature
9	PS00685	1.0	NF-YB/HAP3 subunit signature
9	PS00338	0.8611111111111112	Somatotropin, prolactin and related hormones signature 2
9	PS00937	0.6862745098039215	Ribosomal protein L20 signature
9	PS00968	0.6849845201238389	Antenna complexes alpha subunits signature
9	PS00265	0.580952380952381	Pancreatic hormone family signature

Continued...

#C	#PS	%Helix/Strand	Description
9	PS00331	0.661764705882353	Malic enzymes signature
9	PS00942	0.4117647058823529	glpT family of transporters signature
10	PS01129	0.22222222222222224	Rlu family of pseudouridine synthase signature
10	PS01149	0.05	Rsu family of pseudouridine synthase signature
10	PS01268	0.14285714285714285	Uncharacterized protein family UPF0024 signature
11	PS00109	0.0	Tyrosine protein kinases specific active-site signature
11	PS01245	0.0	RIO1/ZK632.3/MJ0444 family signature
11	PS00108	0.020192307692307697	Serine/Threonine protein kinases active-site signature
12	PS00977	0.6111111111111112	FAD-dependent glycerol-3-phosphate dehydrogenase signature 1
12	PS01238	0.8125	GDA1/CD39 family of nucleoside phosphatases signature
12	PS00712	0.5625	Ribosomal protein S17e signature
13	PS00084	0.625	Copper type II, ascorbate-dependent monooxygenases signature 1
13	PS00289	0.875	Pentaxin family signature
13	PS00319	0.7142857142857143	Amyloidogenic glycoprotein extracellular domain signature
13	PS01295	0.4583333333333333	4-diphosphocytidyl-2C-methyl-D-erythritol synthase signature
13	PS00772	0.0	Barwin domain signature 2
14	PS00034	0.607843137254902	Paired domain signature
14	PS00846	0.6842105263157894	Bacterial regulatory proteins, arsR family signature
14	PS00526	0.15	Ribosomal protein L19e signature
14	PS00356	0.6008771929824561	LacI-type HTH domain signature
15	PS00437	0.5694444444444444	Catalase proximal heme-ligand signature
15	PS01337	0.7777777777777778	Ornithine decarboxylase antizyme signature
16	PS01054	0.6666666666666666	Transaldolase signature 1
16	PS00324	0.4444444444444444	Aspartokinase signature
16	PS00397	0.0	Site-specific recombinases active site
16	PS01048	0.0	Ribosomal protein S6 signature

Continued...

#C	#PS	%Helix/Strand	Description
17	PS00065	0.5642857142857143	D-isomer specific 2-hydroxyacid dehydrogenases NAD-binding signature
17	PS00837	0.564102564102564	Alanine dehydrogenase & pyridine nucleotide transhydrogenase signature 2
18	PS00394	0.6923076923076923	DNA photolyases class 1 signature 1
18	PS00955	0.6923076923076922	Imidazoleglycerol-phosphate dehydratase signature 2
18	PS00271	0.6122448979591837	Plant thionins signature
19	PS00656	0.22499999999999998	Glycosyl hydrolases family 6 signature 2
19	PS01027	0.3	Glycosyl hydrolases family 39 active site
19	PS00659	0.16666666666666666	Glycosyl hydrolases family 5 signature
20	PS01019	0.49094202898550726	ADP-ribosylation factors family signature
20	PS01020	0.4158974358974359	SAR1 family signature
21	PS00185	0.5	Isopenicillin N synthetase signature 1
21	PS00157	0.0	Ribulose bisphosphate carboxylase large chain active site
21	PS00975	0.0	Myristoyl-CoA:protein N-myristoyltransferase signature 1
21	PS00545	0.3	Aldose 1-epimerase putative active site
21	PS00205	0.03125	Transferrins signature 1
22	PS00809	0.2222222222222222	ADP-glucose pyrophosphorylase signature 2
22	PS00177	0.4444444444444444	DNA topoisomerase II signature
22	PS00923	0.2222222222222222	Aspartate and glutamate racemases signature 1
23	PS00120	0.35555555555555557	Lipases, serine active site
23	PS00726	0.3	AP endonucleases family 1 signature 1
24	PS01075	0.5	Acetate and butyrate kinases family signature 1
24	PS00989	0.45454545454545453	Clathrin adaptor complexes small chain signature
25	PS00658	0.5178571428571428	Fork head domain signature 2
25	PS00069	0.0	Glucose-6-phosphate dehydrogenase active site
25	PS00374	0.0	Methylated-DNA–protein-cysteine methyltransferase active site
26	PS00684	0.8666666666666667	Small, acid-soluble spore proteins, alpha/beta type, signature 2

Continued...

#C	#PS	%Helix/Strand	Description
26	PS00511	0.625	Corticotropin-releasing factor family signature
26	PS01350	0.6770833333333334	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase signature
26	PS01259	0.7457142857142857	Apoptosis regulator, Bcl-2 family BH3 motif signature
26	PS00630	0.47878787878787876	Inositol monophosphatase family signature 2
27	PS00763	0.6071428571428571	Glutathione peroxidases signature 2
27	PS00730	0.5	AP endonucleases family 2 signature 2
27	PS00687	0.05555555555555555	Aldehyde dehydrogenases glutamic acid active site
28	PS00010	0.3888888888888888	Aspartic acid and asparagine hydroxylation site
28	PS01336	0.5	S-adenosylmethionine decarboxylase signature
28	PS00626	0.6136363636363635	Regulator of chromosome condensation (RCC1) signature 2
29	PS00123	0.6666666666666666	Alkaline phosphatase active site
29	PS00432	0.3333333333333333	Actins signature 2
29	PS00508	0.1875	Nickel-dependent hydrogenases large subunit signature 2
30	PS01302	0.5	DNA repair protein radC family signature
30	PS00125	0.0	Serine/threonine specific protein phosphatases signature
30	PS00093	0.0	N-4 cytosine-specific DNA methylases signature
30	PS00134	0.032467532467532464	Serine proteases, trypsin family, histidine active site
31	PS00081	0.41818181818181815	Lipoxygenases iron-binding region signature 2
31	PS01009	0.45454545454545453	CRISP family signature 1
32	PS00519	0.5833333333333333	AsnC-type HTH domain signature
32	PS00622	0.6428571428571428	LuxR-type HTH domain signature
32	PS00042	0.6444444444444444	Crp-type HTH domain signature
33	PS00135	0.17	Serine proteases, trypsin family, serine active site
33	PS00673	0.2727272727272727	Serine proteases, V8 family, serine active site

Continued...

#C	#PS	%Helix/Strand	Description
34	PS00465	0.5	POU-specific (POUs) domain signature 2
34	PS00359	0.384478021978022	Ribosomal protein L11 signature
35	PS00092	0.29670329670329665	N-6 Adenine-specific DNA methylases signature
35	PS00485	0.19047619047619047	Adenosine and AMP deaminase signature
35	PS00261	0.0	Glycoprotein hormones beta chain signature 1
35	PS00976	0.0	Myristoyl-CoA:protein N-myristoyltransferase signature 2
36	PS00722	0.8333333333333334	Formate–tetrahydrofolate ligase signature 2
36	PS01265	0.23333333333333334	Tpx family signature
36	PS00978	0.36363636363636365	FAD-dependent glycerol-3-phosphate dehydrogenase signature 2
36	PS00080	0.34523809523809523	Multicopper oxidases signature 2
37	PS00239	0.0	Receptor tyrosine kinase class II signature
37	PS00298	0.3	Heat shock hsp90 proteins family signature
38	PS00149	0.2727272727272727	Sulfatases signature 2
38	PS01122	0.4404761904761905	Caspase family cysteine active site
39	PS01144	0.2	Ribosomal protein L31e signature
39	PS00186	0.5	Isopenicillin N synthetase signature 2
40	PS00152	0.0	ATP synthase alpha and beta subunits signature
40	PS00171	0.14141414141414144	Triosephosphate isomerase active site
41	PS00469	0.0	Nucleoside diphosphate kinases active site
41	PS01095	0.41666666666666663	Chitinases family 18 active site
41	PS00771	0.6666666666666666	Barwin domain signature 1
41	PS01032	0.6111111111111112	Protein phosphatase 2C signature
41	PS00159	0.4000000000000001	KDPG and KHG aldolases active site
41	PS00274	0.7	Aerolysin type toxins signature
42	PS00782	0.8125	Transcription factor TFIIB repeat signature
42	PS01290	0.47058823529411764	Enhancer of rudimentary signature
42	PS00489	0.4	Bacteriophage-type RNA polymerase family active site signature 2
42	PS00788	0.5382352941176471	Chorismate synthase signature 2
43	PS00130	0.3714285714285714	Uracil-DNA glycosylase signature
43	PS00917	0.3636363636363637	Asparaginase / glutaminase active site signature 2
44	PS00184	0.0	Phosphoribosylglycinamide synthetase signature

Continued...

#C	#PS	%Helix/Strand	Description
44	PS60016	0.0	Omega-atracotoxin (ACTX) type 1 family signature
44	PS00336	0.0	Beta-lactamase class-C active site
45	PS00198	0.16176470588235295	4Fe-4S ferredoxin-type iron-sulfur binding region signature
45	PS00276	0.5333333333333334	Channel forming colicins signature
45	PS00102	0.46153846153846156	Phosphorylase pyridoxal-phosphate attachment site
46	PS01044	0.6875	Squalene and phytoene synthases signature 1
46	PS00509	0.39999999999999997	Ras GTPase-activating proteins domain profile
47	PS01070	0.0	DNA/RNA non-specific endonucleases active site
47	PS60009	0.0	Cyclotides Moebius subfamily signature
48	PS00961	0.5555555555555556	Ribosomal protein S28e signature
48	PS00702	0.0	Granulocyte-macrophage colony-stimulating factor signature
48	PS00694	0.0	Enterobacterial virulence outer membrane protein signature 1
48	PS01062	0.0	Hydroxymethylglutaryl-coenzyme A lyase active site
49	PS00155	0.38461538461538464	Cutinase, serine active site
49	PS00411	0.5325757575757575	Kinesin motor domain signature
50	PS01322	0.0	Phosphotriesterase family signature 1
50	PS00695	0.5714285714285714	Enterobacterial virulence outer membrane protein signature 2
50	PS00804	0.0	Calreticulin family signature 2
51	PS00482	0.0	Dihydroorotase signature 1
51	PS01058	0.5833333333333334	SAICAR synthetase signature 2
51	PS00308	0.0	Legume lectins alpha-chain signature
52	PS00129	0.375	Glycosyl hydrolases family 31 active site
52	PS00320	0.0	Amyloidogenic glycoprotein intracellular domain signature
52	PS00172	0.375	Xylose isomerase signature 1
53	PS00046	0.0	Histone H2A signature
53	PS00058	0.0	DNA mismatch repair proteins mutL / hexB / PMS1 signature

Continued...

#C	#PS	%Helix/Strand	Description
54	PS00758	0.45	ArgE / dapE / ACY1 / CPG2 / yscS family signature 1
54	PS00277	0.0	Staphylococcal enterotoxin/Streptococcal pyrogenic exotoxin signature 1
54	PS00591	0.1477272727272727	Glycosyl hydrolases family 10 active site
55	PS00151	0.2450980392156863	Acylphosphatase signature 2
55	PS00678	0.5405982905982906	Trp-Asp (WD) repeats signature
56	PS00631	0.0	Cytosol aminopeptidase signature
56	PS00501	0.0	Signal peptidases I serine active site
56	PS60001	0.0	Nitric oxide synthase (NOS) signature
57	PS00584	0.673469387755102	pfkB family of carbohydrate kinases signature 2
57	PS00258	0.8571428571428571	Calcitonin / CGRP / IAPP family signature
58	PS00928	0.0	Trehalase signature 2
58	PS00197	0.0	2Fe-2S ferredoxin-type iron-sulfur binding region signature
59	PS01090	0.2727272727272727	TatD deoxyribonuclease family signature 2
59	PS00639	0.037037037037037035	Eukaryotic thiol (cysteine) proteases histidine active site
60	PS00566	0.488095238095238	Fibrillarin signature
60	PS00398	0.23076923076923078	Site-specific recombinases signature 2
61	PS00503	0.0	Pectinesterase signature 2
61	PS00410	0.0	Dynamin family signature
62	PS01064	0.5285714285714286	Pyridoxamine 5'-phosphate oxidase signature
62	PS00099	0.5408163265306122	Thiolases active site
62	PS01166	0.34615384615384615	RNA polymerases beta chain signature
63	PS00905	0.35714285714285715	GTP1/OBG family signature
63	PS01201	0.42857142857142855	Tub family signature 2
64	PS00569	0.2857142857142857	Myelin basic protein signature
64	PS00290	0.7802812400127835	Immunoglobulins and major histocompatibility complex proteins signature
65	PS00244	0.7098765432098766	Photosynthetic reaction center proteins signature
65	PS00538	0.3333333333333333	Bacterial chemotaxis sensory transducers signature
66	PS01014	0.35	Transcription termination factor nusG signature

Continued...

#C	#PS	%Helix/Strand	Description
66	PS01097	0.0	Hydrogenases expression/synthesis hupF/hypC family signature
66	PS00710	0.3	Phosphoglucomutase and phosphomannomutase phosphoserine signature
67	PS00063	0.5208333333333334	Aldo/keto reductase family putative active site signature
67	PS01232	0.625	Purine and other phosphorylases family 1 signature
67	PS00201	0.6823529411764706	Flavodoxin signature
68	PS00160	0.0	KDPG and KHG aldolases Schiff-base forming residue
68	PS00169	0.13846153846153847	Delta-aminolevulinic acid dehydratase active site
69	PS00535	0.25	Respiratory chain NADH dehydrogenase 49 Kd subunit signature
69	PS00393	0.5	Phosphoenolpyruvate carboxylase active site 2
69	PS01331	0.3186813186813187	Thymidylate kinase signature
70	PS00318	0.0	Hydroxymethylglutaryl-coenzyme A reductases signature 2
70	PS01266	0.0	Adenylosuccinate synthetase GTP-binding site
71	PS00279	0.5	Membrane attack complex components / perforin signature
71	PS01306	0.25	Uncharacterized protein family UPF0054 signature
72	PS01164	0.5476190476190476	Copper amine oxidase topaquinone signature
72	PS01226	0.0	Hydroxymethylglutaryl-coenzyme A synthase active site
73	PS01196	0.5	Peptidyl-tRNA hydrolase signature 2
73	PS01247	0.4318181818181818	Inosine-uridine preferring nucleoside hydrolase family signature
74	PS01224	0.5058823529411764	N-acetyl-gamma-glutamyl-phosphate reductase active site

Continued...

#C	#PS	%Helix/Strand	Description
74	PS00778	0.680672268907563	Histidine acid phosphatases active site signature
75	PS00147	0.4166666666666666	Arginase family signature 1
75	PS00168	0.43333333333333335	Tryptophan synthase beta chain pyridoxal-phosphate attachment site
76	PS00776	0.45454545454545453	Glycosyl hydrolases family 11 active site signature 1
76	PS01089	0.5	CAP protein signature 2
76	PS01153	0.25	NOL1/NOP2/sun family signature
76	PS01172	0.3333333333333333	Ribosomal protein L44e signature
77	PS00506	0.0	Beta-amylase active site 1
77	PS00572	0.2962962962962963	Glycosyl hydrolases family 1 active site
78	PS00915	0.4666666666666667	Phosphatidylinositol 3- and 4-kinases signature 1
78	PS00938	0.6428571428571429	Initiation factor 3 signature
79	PS00144	0.3385416666666667	Asparaginase / glutaminase active site signature 1
79	PS00697	0.0	ATP-dependent DNA ligase AMP-binding site
79	PS00148	0.3333333333333333	Arginase family signature 2
80	PS00321	0.14814814814814814	recA signature
80	PS00173	0.04285714285714286	Xylose isomerase signature 2
81	PS00742	0.47368421052631576	PEP-utilizing enzymes signature 2
81	PS00907	0.6470588235294118	Uroporphyrinogen decarboxylase signature 2
82	PS00088	0.0	Manganese and iron superoxide dismutases signature
82	PS00369	0.30357142857142855	PTS HPR domain histidine phosphorylation site signature
83	PS00100	0.0	Chloramphenicol acetyltransferase active site
83	PS00158	0.2727272727272727	Fructose-bisphosphate aldolase class-I active site
84	PS00644	0.4375	Respiratory-chain NADH dehydrogenase 51 Kd subunit signature 1
84	PS01036	0.43333333333333335	Heat shock hsp70 proteins family signature 3
85	PS01317	0.5384615384615385	SsrA-binding protein
85	PS00288	0.2727272727272727	Tissue inhibitors of metalloproteinases signature

Continued...

#C	#PS	%Helix/Strand	Description
86	PS00833	0.0	2'-5'-oligoadenylate synthetases signature 2
86	PS01034	0.5426136363636364	Glycosyl hydrolases family 16 active sites
87	PS00032	0.0	'Homeobox' antennapedia-type protein signature
87	PS00341	0.0	Surfactant associated polypeptide SP-C palmitoylation sites
88	PS00381	0.25	Endopeptidase Clp serine active site
88	PS00956	0.3333333333333333	Fungal hydrophobins signature
89	PS00902	0.5714285714285714	Glutamate 5-kinase signature
89	PS00146	0.5	Beta-lactamase class-A active site
90	PS00842	0.6	XPG protein signature 2
90	PS01057	0.3333333333333333	SAICAR synthetase signature 1
91	PS00087	0.0	Copper/Zinc superoxide dismutase signature 1
91	PS01010	0.48611111111111116	CRISP family signature 2
92	PS00699	0.0	Nitrogenases component 1 alpha and beta subunits signature 1
92	PS00546	0.0	Matrixins cysteine switch
93	PS00055	0.0	Ribosomal protein S12 signature
93	PS00097	0.0	Aspartate and ornithine carbamoyltransferases signature
94	PS00059	0.0	Zinc-containing alcohol dehydrogenases signature
94	PS01334	0.0	Pyrrolidone-carboxylate peptidase cysteine active site
95	PS00354	0.0	HMG-I and HMG-Y DNA-binding domain (A+T-hook)
95	PS01236	0.34545454545454546	PdxT/SNO family family signature
96	PS00231	0.0	F-actin capping protein beta subunit signature
96	PS00496	0.0	P-II protein uridylation site
97	PS01231	0.0	RNA methyltransferase trmA family signature 2
97	PS01013	0.2727272727272727	Oxysterol-binding protein family signature
98	PS00094	0.07692307692307693	C-5 cytosine-specific DNA methylases active site
98	PS01158	0.3833333333333333	Macrophage migration inhibitory factor family signature

Continued...

#C	#PS	%Helix/Strand	Description
99	PS00131	0.0	Serine carboxypeptidases, serine active site
99	PS00297	0.375	Heat shock hsp70 proteins family signature 1
100	PS00074	0.2976190476190476	Glu / Leu / Phe / Val dehydrogenases active site
100	PS00085	0.23076923076923078	Copper type II, ascorbate-dependent monooxygenases signature 2
100	PS01121	0.3333333333333333	Caspase family histidine active site
101	PS00071	0.5882352941176471	Glyceraldehyde 3-phosphate dehydrogenase active site
101	PS00414	0.5833333333333333	Profilin signature
102	PS01335	0.3333333333333333	Methylglyoxal synthase active site
102	PS00536	0.3333333333333333	Ubiquitin-activating enzyme signature 1
103	PS00027	0.5913947163947164	'Homeobox' domain signature
103	PS00278	0.49537037037037035	Staphyloccocal enterotoxin/Streptococcal pyrogenic exotoxin signature 2
104	PS00653	0.14666666666666667	Glycosyl hydrolases family 1 N-terminal signature
104	PS00868	0.044444444444444446	Cys/Met metabolism enzymes pyridoxal-phosphate attachment site
105	PS00589	0.15625	PTS HPR domain serine phosphorylation site signature
105	PS00145	0.3137254901960784	Urease active site
106	PS00497	0.5972222222222222	Tyrosinase CuA-binding region signature
106	PS00098	0.5789473684210527	Thiolases acyl-enzyme intermediate signature
107	PS00732	0.0	Ribosomal protein S16 signature
107	PS00111	0.3305785123966943	Phosphoglycerate kinase signature
108	PS00727	0.17647058823529413	AP endonucleases family 1 signature 2
108	PS00896	0.2	LacY family proton/sugar symporters signature 1
109	PS01274	0.3333333333333333	Coenzyme A transferases signature 2
109	PS00564	0.3333333333333333	Argininosuccinate synthase signature 1
110	PS01137	0.16666666666666666	TatD deoxyribonuclease family signature 1
110	PS00175	0.03333333333333333	Phosphoglycerate mutase family phosphohistidine signature
111	PS00850	0.0	Glycine radical domain signature
111	PS00419	0.0	Photosystem I psaA and psaB proteins signature

Continued...

#C	#PS	%Helix/Strand	Description
112	PS01181	0.0	Ribosomal protein S21 signature
112	PS00424	0.0	Interleukin-2 signature
113	PS00096	0.0	Serine hydroxymethyltransferase pyridoxal-phosphate attachment site
113	PS00853	0.0	Beta-eliminating lyases pyridoxal-phosphate attachment site
114	PS00691	0.5125	DNA photolyases class 1 signature 2
114	PS01026	0.16666666666666666	Photosystem I psaG and psaK proteins signature
115	PS00154	0.0	E1-E2 ATPases phosphorylation site
115	PS00227	0.0	Tubulin subunits alpha, beta, and gamma signature
116	PS00122	0.18303571428571427	Carboxylesterases type-B serine active site
116	PS01091	0.23529411764705885	TatD deoxyribonuclease family signature 3
117	PS01284	0.0	Thermonuclease family signature 2
117	PS00133	0.0	Zinc carboxypeptidases, zinc-binding region 2 signature
118	PS00704	0.0	Prokaryotic-type carbonic anhydrases signature 1
118	PS00242	0.0	Integrins alpha chain signature
119	PS00944	0.43421052631578944	Cyclin-dependent kinases regulatory subunits signature 1
119	PS01203	0.45	BTG family signature 2
120	PS00530	0.0	Ribonuclease T2 family histidine active site 1
120	PS00867	0.0	Carbamoyl-phosphate synthase subdomain signature 2
121	PS00067	0.54	3-hydroxyacyl-CoA dehydrogenase signature
121	PS00254	0.6346153846153846	Interleukin-6 / G-CSF / MGF signature
122	PS01016	0.42857142857142855	Glycoprotease family signature
122	PS01139	0.36250000000000004	Bacterial microcompartiments proteins signature
123	PS00671	0.2398190045248869	D-isomer specific 2-hydroxyacid dehydrogenases signature 3
123	PS01103	0.33333333333333337	Aspartate-semialdehyde dehydrogenase signature
124	PS00683	0.4696969696969697	Rhodanese C-terminal signature
124	PS00906	0.0	Uroporphyrinogen decarboxylase signature 1

Continued...

#C	#PS	%Helix/Strand	Description
125	PS00859	0.2647058823529412	GTP cyclohydrolase I signature 1
125	PS01156	0.5146103896103896	TonB-dependent receptor proteins signature 2
126	PS00794	0.38888888888888884	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase signature
126	PS00136	0.2588383838383838	Serine proteases, subtilase family, aspartic acid active site
127	PS00616	0.3	Histidine acid phosphatases phosphohistidine signature
127	PS00587	0.2857142857142857	Glycosyl hydrolases family 17 signature
128	PS00577	0.28775510204081634	Avidin-like domain signature
128	PS00935	0.08823529411764706	Glyoxalase I signature 2
129	PS00602	0.26785714285714285	Fructose-bisphosphate aldolase class-II signature 1
129	PS00401	0.5384615384615384	Prokaryotic sulfate-binding proteins signature 1
130	PS00679	0.0	Beta-amylase active site 2
130	PS01151	0.2727272727272727	Fimbrial biogenesis outer membrane usher protein signature
131	PS00194	0.4978070175438596	Thioredoxin family active site
131	PS01135	0.356060606060606	FtsZ protein signature 2
132	PS00916	0.5238095238095238	Phosphatidylinositol 3- and 4-kinases signature 2
132	PS00447	0.7250000000000001	DNA polymerase family A signature
133	PS00086	0.0	Cytochrome P450 cysteine heme-iron ligand signature
133	PS00116	0.24999999999999997	DNA polymerase family B signature
134	PS00924	0.23376623376623376	Aspartate and glutamate racemases signature 2
134	PS01127	0.6060606060606061	Elongation factor Ts signature 2
135	PS00743	0.25595238095238093	Beta-lactamases class B signature 1
135	PS00209	0.575	Arthropod hemocyanins / insect LSPs signature 1
136	PS00072	0.0	Acyl-CoA dehydrogenases signature 1
136	PS00480	0.0	Citrate synthase signature
137	PS00505	0.2222222222222222	Phosphoenolpyruvate carboxykinase (GTP) signature
137	PS00221	0.6296296296296295	MIP family signature

Continued...

#C	#PS	%Helix/Strand	Description
138	PS01008	0.625	DnaA protein signature
138	PS00878	0.4511278195488721	Orn/DAP/Arg decarboxylases family 2 pyridoxal-P attachment site
139	PS00371	0.38461538461538464	PTS EIIA domains phosphorylation site signature 1
139	PS00900	0.6666666666666666	Bacteriophage-type RNA polymerase family active site signature 1
140	PS00406	0.3636363636363636	Actins signature 1
140	PS00335	0.13333333333333333	Parathyroid hormone family signature
141	PS00761	0.21428571428571427	Signal peptidases I signature 3
141	PS01012	0.25	Folylpolyglutamate synthase signature 2
142	PS00024	0.026785714285714284	Hemopexin domain signature
142	PS01015	0.0	Ribosomal protein L19 signature
143	PS00035	0.46153846153846156	POU-specific (POUs) domain signature 1
143	PS00642	0.0	Respiratory-chain NADH dehydrogenase 75 Kd subunit signature 2
144	PS00139	0.3906249999999999	Eukaryotic thiol (cysteine) proteases cysteine active site
144	PS00860	0.45454545454545453	GTP cyclohydrolase I signature 2
145	PS00332	0.07272727272727272	Copper/Zinc superoxide dismutase signature 2
145	PS01305	0.0	moaA / nifB / pqqE family signature
146	PS00625	0.0	Regulator of chromosome condensation (RCC1) signature 1
146	PS00416	0.1818181818181818	Synapsins signature 2
147	PS00518	0.31818181818181823	Zinc finger RING-type signature
147	PS00721	0.0	Formate–tetrahydrofolate ligase signature 1
148	PS00012	0.5480769230769231	Phosphopantetheine attachment site
149	PS00888	0.15032679738562094	Cyclic nucleotide-binding domain signature 1
150	PS00889	0.06349206349206349	Cyclic nucleotide-binding domain signature 2
151	PS00020	0.584	Actinin-type actin-binding domain signature 2
152	PS01177	0.34841269841269845	Anaphylatoxin domain signature
153	PS00495	0.2155197444831591	Apple domain
154	PS60024	0.036585365853658534	Agouti domain signature
155	PS00427	0.0	Disintegrins signature
156	PS00018	0.187823834196891	EF-hand calcium-binding domain

Continued...

#C	#PS	%Helix/Strand	Description
157	PS00022	0.03571428571428571	EGF-like domain signature 1
158	PS01186	0.023697916666666666	EGF-like domain signature 2
159	PS01187	0.16693964930259825	Calcium-binding EGF-like domain signature
160	PS01248	0.0	Laminin-type EGF-like (LE) domain signature
161	PS01285	0.2593360071301248	Coagulation factors 5/8 type C domain (FA58C) signature 1
162	PS01286	0.18823529411764708	Coagulation factors 5/8 type C domain (FA58C) signature 2
163	PS00660	0.10474601408972932	FERM domain signature 1
164	PS00661	0.35	FERM domain signature 2
165	PS01253	0.2797292376239745	Fibronectin type-I domain signature
166	PS00023	0.005952380952380952	Fibronectin type-II collagen-binding domain signature
167	PS00514	0.0	Fibrinogen beta and gamma chains C-terminal domain signature
168	PS00011	0.0	Vitamin K-dependent carboxylation domain
169	PS00222	0.0	Insulin-like growth factor-binding protein (IGFBP) N-terminal domain signature
170	PS00021	0.0	Kringle domain signature
171	PS01209	0.017777777777777778	LDL-receptor class A (LDLRA) domain signature
172	PS00615	0.21784261193352103	C-type lectin domain signature
173	PS00478	0.02288778759366995	LIM zinc-binding domain signature
174	PS01241	0.2351046698872786	Link domain signature
175	PS00612	0.0	Osteonectin domain signature 1
176	PS00025	0.1593073593073593	P-type 'Trefoil' domain signature
177	PS00562	0.0	CBM1 (carbohydrate binding type-1) domain signature
178	PS00561	0.0	CBM2a (carbohydrate-binding type-2) domain signature
179	PS00026	0.0	Chitin recognition or binding domain signature
180	PS01282	0.22432512565577767	BIR repeat
181	PS00845	0.1599264705882353	CAP-Gly domain signature
182	PS00983	0.20388771435283065	Ly-6 / u-PAR domain signature
183	PS00740	0.525	MAM domain signature
184	PS00524	0.0	Somatomedin B domain (SMB) signature
185	PS00420	0.34210526315789475	SRCR domain signature

Continued...

#C	#PS	%Helix/Strand	Description
186	PS00484	0.1227450980392157	Thyroglobulin type-1 repeat signature
187	PS01208	0.07894736842105263	VWFC domain signature
188	PS01159	0.2306742640075973	WW/rsp5/WWP domain signature
189	PS01049	0.0	YjeF C-terminal domain signature 1
190	PS01360	0.10855855855855856	Zinc finger MYND-type signature
191	PS01359	0.09365882891748267	Zinc finger PHD-type signature
192	PS00479	0.09486297372895765	Zinc finger phorbol-ester/DAG-type signature
193	PS01358	0.08333333333333333	Zinc finger RanBP2-type signature
194	PS01357	0.10661113080467918	Zinc finger ZZ-type signature
195	PS00028	0.2864051930832814	Zinc finger C2H2 type domain signature
196	PS00466	0.22222222222222224	Zinc finger TFIIS-type signature
197	PS00031	0.27697649572649574	Nuclear hormones receptors DNA-binding region signature
198	PS00344	0.09333333333333334	GATA-type zinc finger domain
199	PS00347	0.18615984405458086	Poly(ADP-ribose) polymerase zinc finger domain signature
200	PS00463	0.2652167374179928	Zn(2)-C6 fungal-type DNA-binding domain signature
201	PS01102	0.2	Prokaryotic dksA C4-type zinc finger
202	PS01119	0.10810810810810811	Copper-fist DNA-binding domain signature
203	PS00348	0.0	p53 family signature
204	PS00352	0.2628654970760234	'Cold-shock' domain signature
205	PS40000	0.0	DM DNA-binding domain signature
206	PS00346	0.4140625	Ets-domain signature 2
207	PS00657	0.49313186813186816	Fork head domain signature 1
208	PS00434	0.33999999999999997	HSF-type DNA-binding domain signature
209	PS00601	0.19986631016042783	Tryptophan pentad repeat (IRF family) signature
210	PS01204	0.0	NF-kappa-B/Rel/dorsal domain signature
211	PS00350	0.5780303030303031	MADS-box domain signature
212	PS01283	0.6	T-box domain signature 1
213	PS01264	0.18421052631578946	T-box domain signature 2
214	PS00554	0.5172413793103449	TEA domain signature
215	PS00351	0.472	Transcription factor TFIID repeat signature
216	PS01289	0.0	TSC-22 / dip / bun family signature
217	PS00829	0.5	Prokaryotic transcription elongation factors signature 1
218	PS00830	0.23529411764705882	Prokaryotic transcription elongation factors signature 2

Continued...

#C	#PS	%Helix/Strand	Description
219	PS00039	0.40740740740740744	DEAD-box subfamily ATP-dependent helicases signature
220	PS00752	0.0	XPA protein signature 1
221	PS00753	0.0	XPA protein signature 2
222	PS00841	0.39999999999999997	XPG protein signature 1
223	PS00041	0.59824738793394	Bacterial regulatory proteins, araC family signature
224	PS01117	0.4380952380952381	MarR-type HTH domain signature
225	PS00552	0.5942028985507246	MerR-type HTH domain signature
226	PS01081	0.4526209677419355	TetR-type HTH domain signature
227	PS00716	0.7222222222222222	Sigma-70 factors family signature 2
228	PS01063	0.65625	Sigma-70 factors ECF subfamily signature
229	PS00675	0.42857142857142855	Sigma-54 interaction domain ATP-binding region A signature
230	PS00676	0.4375	Sigma-54 interaction domain ATP-binding region B signature
231	PS00688	0.5	Sigma-54 interaction domain C-terminal part signature
232	PS00045	0.20634920634920637	Bacterial histone-like DNA-binding proteins signature
233	PS00818	0.4117647058823529	Dps protein family signature 1
234	PS00617	0.3333333333333333	RecF protein signature 1
235	PS00618	0.47368421052631576	RecF protein signature 2
236	PS01300	0.0	RecR protein signature
237	PS00357	0.7681159420289855	Histone H2B signature
238	PS00322	0.0	Histone H3 signature 1
239	PS00633	0.37231422913034656	Bromodomain signature
240	PS00598	0.1365646258503401	Chromo domain signature
241	PS01199	0.24298245614035088	Ribosomal protein L1 signature
242	PS00467	0.0	Ribosomal protein L2 signature
243	PS00474	0.08333333333333333	Ribosomal protein L3 signature
244	PS00358	0.27941176470588236	Ribosomal protein L5 signature
245	PS00700	0.15	Ribosomal protein L6 signature 2
246	PS00651	0.0	Ribosomal protein L9 signature
247	PS01109	0.6571428571428571	Ribosomal protein L10 signature
248	PS00783	0.17889492753623187	Ribosomal protein L13 signature
249	PS00049	0.15740740740740738	Ribosomal protein L14 signature
250	PS00475	0.03125	Ribosomal protein L15 signature
251	PS00586	0.3333333333333333	Ribosomal protein L16 signature 1
252	PS00701	0.0	Ribosomal protein L16 signature 2
253	PS01169	0.5	Ribosomal protein L21 signature
254	PS00464	0.5066666666666667	Ribosomal protein L22 signature

Continued...

#C	#PS	%Helix/Strand	Description
255	PS00050	0.3333333333333333	Ribosomal protein L23 signature
256	PS01108	0.12962962962962962	Ribosomal protein L24 signature
257	PS00831	0.0	Ribosomal protein L27 signature
258	PS00579	0.4666666666666666	Ribosomal protein L29 signature
259	PS00634	0.3787878787878788	Ribosomal protein L30 signature
260	PS01143	0.20454545454545453	Ribosomal protein L31 signature
261	PS00582	0.125	Ribosomal protein L33 signature
262	PS00784	0.07894736842105263	Ribosomal protein L34 signature
263	PS00936	0.14814814814814814	Ribosomal protein L35 signature
264	PS00828	0.32	Ribosomal protein L36 signature
265	PS00939	0.5925925925925926	Ribosomal protein L1e signature
266	PS01082	0.5952380952380952	Ribosomal protein L7Ae signature
267	PS01257	0.0	Ribosomal protein L10e signature
268	PS01194	0.47222222222222215	Ribosomal protein L15e signature
269	PS01106	0.16666666666666666	Ribosomal protein L18e signature
270	PS01171	0.11538461538461539	Ribosomal protein L21e signature
271	PS01073	0.2222222222222222	Ribosomal protein L24e signature
272	PS00709	0.4828571428571429	Ribosomal protein L30e signature 1
273	PS00993	0.22619047619047616	Ribosomal protein L30e signature 2
274	PS00580	0.0	Ribosomal protein L32e signature
275	PS01105	0.13636363636363635	Ribosomal protein L35Ae signature
276	PS01077	0.0	Ribosomal protein L37e signature
277	PS00962	0.1875	Ribosomal protein S2 signature 1
278	PS00963	0.38	Ribosomal protein S2 signature 2
279	PS00548	0.4123552123552124	Ribosomal protein S3 signature
280	PS00632	0.54	Ribosomal protein S4 signature
281	PS00585	0.6515151515151515	Ribosomal protein S5 signature
282	PS00052	0.5308641975308642	Ribosomal protein S7 signature
283	PS00053	0.4222222222222222	Ribosomal protein S8 signature
284	PS00361	0.0625	Ribosomal protein S10 signature
285	PS00054	0.2717391304347826	Ribosomal protein S11 signature
286	PS00646	0.3214285714285714	Ribosomal protein S13 signature
287	PS00527	0.17654808959156787	Ribosomal protein S14 signature
288	PS00362	0.4838709677419355	Ribosomal protein S15 signature
289	PS00056	0.0	Ribosomal protein S17 signature
290	PS00057	0.0	Ribosomal protein S18 signature
291	PS00323	0.1	Ribosomal protein S19 signature
292	PS00529	0.40217391304347827	Ribosomal protein S24e signature
293	PS01168	0.09090909090909091	Ribosomal protein S27e signature
294	PS00486	0.3529411764705882	DNA mismatch repair proteins mutS family signature
295	PS00893	0.3896103896103896	Nudix hydrolase signature
296	PS01162	0.5568181818181818	Quinone oxidoreductase / zeta-crystallin signature

Continued...

#C	#PS	%Helix/Strand	Description
297	PS00913	0.6827586206896552	Iron-containing alcohol dehydrogenases signature 1
298	PS00060	0.5238095238095237	Iron-containing alcohol dehydrogenases signature 2
299	PS00061	0.5686407221765878	Short-chain dehydrogenases/reductases family signature
300	PS00798	0.33333333333333337	Aldo/keto reductase family signature 1
301	PS00062	0.31196581196581197	Aldo/keto reductase family signature 2
302	PS01042	0.34782608695652173	Homoserine dehydrogenase signature
303	PS00611	0.5151515151515151	Histidinol dehydrogenase signature
304	PS00064	0.0	L-lactate dehydrogenase active site
305	PS00670	0.16304347826086954	D-isomer specific 2-hydroxyacid dehydrogenases signature 2
306	PS00895	0.5714285714285714	3-hydroxyisobutyrate dehydrogenase signature
307	PS00066	0.30000000000000004	Hydroxymethylglutaryl-coenzyme A reductases signature 1
308	PS01192	0.41483516483516486	Hydroxymethylglutaryl-coenzyme A reductases signature 3
309	PS00461	0.2884615384615385	6-phosphogluconate dehydrogenase signature
310	PS00487	0.0	IMP dehydrogenase / GMP reductase signature
311	PS00363	0.0	Bacterial quinoprotein dehydrogenases signature 1
312	PS00364	0.20909090909090908	Bacterial quinoprotein dehydrogenases signature 2
313	PS00557	0.0	FMN-dependent alpha-hydroxy acid dehydrogenases active site
314	PS00623	0.14166666666666666	GMC oxidoreductases signature 1
315	PS00559	0.39176245210727967	Eukaryotic molybdopterin oxidoreductases signature
316	PS00551	0.3148148148148148	Prokaryotic molybdopterin oxidoreductases signature 1
317	PS00490	0.3111111111111111	Prokaryotic molybdopterin oxidoreductases signature 2

Continued...

#C	#PS	%Helix/Strand	Description
318	PS00932	0.38265306122448983	Prokaryotic molybdopterin oxidoreductases signature 3
319	PS00070	0.0	Aldehyde dehydrogenases cysteine active site
320	PS01223	0.5	Gamma-glutamyl phosphate reductase signature
321	PS01298	0.4444444444444444	Dihydrodipicolinate reductase signature
322	PS00911	0.0	Dihydroorotate dehydrogenase signature 1
323	PS00912	0.4047619047619047	Dihydroorotate dehydrogenase signature 2
324	PS01021	0.28	Coproporphyrinogen III oxidase signature
325	PS00504	0.0	Fumarate reductase / succinate dehydrogenase FAD-binding site
326	PS00073	0.2944444444444444	Acyl-CoA dehydrogenases signature 2
327	PS00836	0.2962962962962963	Alanine dehydrogenase & pyridine nucleotide transhydrogenase signature 1
328	PS00677	0.3684210526315789	D-amino acid oxidases signature
329	PS01165	0.12857142857142856	Copper amine oxidase copper-binding site signature
330	PS00521	0.6521739130434783	Delta 1-pyrroline-5-carboxylate reductase signature
331	PS00075	0.3687290969899666	Dihydrofolate reductase (DHFR) domain signature
332	PS00766	0.5	Tetrahydrofolate dehydrogenase/cyclohydrolase signature 1
333	PS00862	0.4864864864864865	Oxygen oxidoreductases covalent FAD-binding site
334	PS00076	0.0	Pyridine nucleotide-disulphide oxidoreductases class-I active site
335	PS00573	0.16060606060606059	Pyridine nucleotide-disulphide oxidoreductases class-II active site
336	PS01150	0.0	Respiratory-chain NADH dehydrogenase 20 Kd subunit signature
337	PS01099	0.0	Respiratory-chain NADH dehydrogenase 24 Kd subunit signature
338	PS00542	0.3181818181818182	Respiratory chain NADH dehydrogenase 30 Kd subunit signature

Continued...

#C	#PS	%Helix/Strand	Description
339	PS00645	0.0	Respiratory-chain NADH dehydrogenase 51 Kd subunit signature 2
340	PS00641	0.0	Respiratory-chain NADH dehydrogenase 75 Kd subunit signature 1
341	PS00643	0.0	Respiratory-chain NADH dehydrogenase 75 Kd subunit signature 3
342	PS00365	0.0	Nitrite and sulfite reductases iron-sulfur/siroheme-binding site
343	PS00366	0.21428571428571427	Uricase signature
344	PS00077	0.5361038961038961	Heme-copper oxidase catalytic subunit, copper B binding region signature
345	PS00078	0.29591836734693877	CO II and nitrous oxide reductase dinuclear copper centers signature
346	PS00848	0.41304347826086957	Cytochrome c oxidase subunit Vb, zinc binding region signature
347	PS01329	0.0	Cytochrome c oxidase subunit VIa signature
348	PS00079	0.2773109243697479	Multicopper oxidases signature 1
349	PS00438	0.0	Catalase proximal active site signature
350	PS00460	0.2	Glutathione peroxidases active site
351	PS00711	0.05333333333333333	Lipoxygenases iron-binding region signature 1
352	PS00082	0.13636363636363635	Extradiol ring-cleavage dioxygenases signature
353	PS00083	0.18226600985221678	Intradiol ring-cleavage dioxygenases signature
354	PS00876	0.2727272727272727	Indoleamine 2,3-dioxygenase signature 1
355	PS00877	0.5714285714285714	Indoleamine 2,3-dioxygenase signature 2
356	PS00570	0.125	Bacterial ring hydroxylating dioxygenases alpha-subunit signature
357	PS00494	0.5096153846153846	Bacterial luciferase subunits signature
358	PS00574	0.6	Fatty acid desaturases family 2 signature
359	PS00089	0.3333333333333333	Ribonucleotide reductase large subunit signature
360	PS00090	0.3888888888888889	Nitrogenases component 1 alpha and beta subunits signature 2
361	PS00746	0.0	NifH/frxC family signature 1
362	PS00692	0.10714285714285714	NifH/frxC family signature 2

Continued...

#C	#PS	%Helix/Strand	Description
363	PS00507	0.23076923076923078	Nickel-dependent hydrogenases large subunit signature 1
364	PS00747	0.5	Glutamyl-tRNA reductase signature
365	PS01100	0.1323529411764706	NNMT/PNMT/TEMT family of methyltransferases signature
366	PS01230	0.29516129032258065	RNA methyltransferase trmA family signature 1
367	PS00091	0.20061576354679808	Thymidylate synthase active site
368	PS01131	0.375	Ribosomal RNA adenine dimethylases signature
369	PS00095	0.7543859649122807	C-5 cytosine-specific DNA methylases C-terminal signature
370	PS01279	0.0	Protein-L-isoaspartate(D-aspartate) O-methyltransferase signature
371	PS00839	0.06666666666666667	Uroporphyrin-III C-methyltransferase signature 1
372	PS00840	0.42647058823529416	Uroporphyrin-III C-methyltransferase signature 2
373	PS00373	0.28125	Phosphoribosylglycinamide formyltransferase active site
374	PS00801	0.563095238095238	Transketolase signature 1
375	PS00802	0.2647058823529412	Transketolase signature 2
376	PS00958	0.611111111111111	Transaldolase active site
377	PS00439	0.5	Acyltransferases ChoActase / COT / CPT family signature 1
378	PS00440	0.20238095238095236	Acyltransferases ChoActase / COT / CPT family signature 2
379	PS00737	0.0	Thiolases signature 2
380	PS00101	0.04310344827586207	Hexapeptide-repeat containing-transferases signature
381	PS00606	0.47593582887700536	Beta-ketoacyl synthases active site
382	PS00441	0.5882352941176471	Chalcone and stilbene synthases active site
383	PS00462	0.4565217391304348	Gamma-glutamyltranspeptidase signature
384	PS00375	0.35353535353535354	UDP-glycosyltransferases signature
385	PS00103	0.23626373626373626	Purine/pyrimidine phosphoribosyl transferases signature

Continued...

#C	#PS	%Helix/Strand	Description
386	PS01240	0.374274099883856	Purine and other phosphorylases family 2 signature
387	PS00647	0.20833333333333334	Thymidine and pyrimidine-nucleoside phosphorylases signature
388	PS01316	0.02727272727272727	ATP phosphoribosyltransferase signature
389	PS00376	0.42424242424242425	S-adenosylmethionine synthetase signature 1
390	PS00377	0.0	S-adenosylmethionine synthetase signature 2
391	PS00723	0.37843137254901965	Polyprenyl synthetases signature 1
392	PS01066	0.16666666666666666	Undecaprenyl pyrophosphate synthetase family signature
393	PS01330	0.5	Spermidine/spermine synthases family signature
394	PS01045	0.6538461538461539	Squalene and phytoene synthases signature 2
395	PS00792	0.34375	Dihydropteroate synthase signature 1
396	PS00793	0.3333333333333333	Dihydropteroate synthase signature 2
397	PS00104	0.35555555555555557	EPSP synthase signature 1
398	PS00885	0.47368421052631576	EPSP synthase signature 2
399	PS00105	0.016483516483516484	Aminotransferases class-I pyridoxal-phosphate attachment site
400	PS00599	0.0	Aminotransferases class-II pyridoxal-phosphate attachment site
401	PS00600	0.1208481507823613	Aminotransferases class-III pyridoxal-phosphate attachment site
402	PS00770	0.21269841269841272	Aminotransferases class-IV signature
403	PS00595	0.19047619047619047	Aminotransferases class-V pyridoxal-phosphate attachment site
404	PS00378	0.27884615384615385	Hexokinases signature
405	PS00106	0.27777777777777773	Galactokinase signature
406	PS00627	0.25	GHMP kinases putative ATP-binding domain
407	PS00433	0.42105263157894735	Phosphofructokinase signature
408	PS00583	0.52	pfkB family of carbohydrate kinases signature 1
409	PS01125	0.12362637362637363	ROK family signature
410	PS00567	0.5	Phosphoribulokinase signature
411	PS00603	0.07142857142857142	Thymidine kinase cellular-type signature
412	PS00445	0.44047619047619047	FGGY family of carbohydrate kinases signature 2

Continued...

#C	#PS	%Helix/Strand	Description
413	PS00107	0.19094681011873335	Protein kinases ATP-binding region signature
414	PS01351	0.4064911625591237	MAP kinase signature
415	PS01101	0.09375	Casein kinase II regulatory subunit signature
416	PS00110	0.6538461538461539	Pyruvate kinase active site signature
417	PS01128	0.5088433048433048	Shikimate kinase signature
418	PS01076	0.5416666666666667	Acetate and butyrate kinases family signature 2
419	PS00112	0.0	ATP:guanido phosphotransferases active site
420	PS00372	0.17647058823529413	PTS EIIA domains phosphorylation site signature 2
421	PS01035	0.16666666666666666	PTS EIIB domains cysteine phosphorylation site signature
422	PS00113	0.4375	Adenylate kinase signature
423	PS00856	0.0	Guanylate kinase-like signature
424	PS00114	0.0	Phosphoribosyl pyrophosphate synthetase signature
425	PS01030	0.18518518518518517	RNA polymerases M / 15 Kd subunits signature
426	PS00446	0.45121951219512196	RNA polymerases D / 30 to 40 Kd subunits signature
427	PS01110	0.14285714285714285	RNA polymerases H / 23 Kd subunits signature
428	PS01154	0.3125	RNA polymerases L / 13 to 16 Kd subunits signature
429	PS01112	0.0	RNA polymerases N / 8 Kd subunits signature
430	PS00522	0.27999999999999997	DNA polymerase family X signature
431	PS00117	0.0	Galactose-1-phosphate uridyl transferase family 1 active site signature
432	PS00808	0.0	ADP-glucose pyrophosphorylase signature 1
433	PS00810	0.13636363636363635	ADP-glucose pyrophosphorylase signature 3
434	PS01277	0.4743589743589744	Ribonuclease PH signature
435	PS00370	0.49053030303030304	PEP-utilizing enzymes phosphorylation site signature
436	PS00380	0.08333333333333333	Rhodanese signature 1
437	PS01273	0.22916666666666666	Coenzyme A transferases signature 1
438	PS00118	0.9797297297297297	Phospholipase A2 histidine active site

Continued...

#C	#PS	%Helix/Strand	Description
439	PS00121	0.0	Colipase signature
440	PS01098	0.2727272727272727	Lipolytic enzymes "G-D-S-L" family, serine active site
441	PS00941	0.4304812834224598	Carboxylesterases type-B signature 2
442	PS00800	0.42500000000000004	Pectinesterase signature 1
443	PS01195	0.4047619047619047	Peptidyl-tRNA hydrolase signature 1
444	PS01328	0.0	4-hydroxybenzoyl-CoA thioesterase family active site
445	PS00785	0.46153846153846156	5'-nucleotidase signature 1
446	PS00786	0.0	5'-nucleotidase signature 2
447	PS00124	0.38461538461538464	Fructose-1-6-bisphosphatase active site
448	PS00383	0.00606060606060606	Tyrosine specific protein phosphatases active site
449	PS00629	0.5663919413919413	Inositol monophosphatase family signature 1
450	PS00384	0.6666666666666666	Prokaryotic zinc-dependent phospholipase C signature
451	PS00126	0.041666666666666664	3'5'-cyclic nucleotide phosphodiesterases signature
452	PS00728	0.125	AP endonucleases family 1 signature 3
453	PS00729	0.0	AP endonucleases family 2 signature 1
454	PS00731	0.17647058823529413	AP endonucleases family 2 signature 3
455	PS00919	0.5714285714285714	Deoxyribonuclease I signature 1
456	PS00918	0.0	Deoxyribonuclease I signature 2
457	PS01321	0.42857142857142855	Crossover junction endodeoxyribonuclease ruvC signature
458	PS00764	0.0	Endonuclease III iron-sulfur binding region signature
459	PS01155	0.4000000000000001	Endonuclease III family signature
460	PS01175	0.22	Ribonuclease II family signature
461	PS00127	0.4	Pancreatic ribonuclease family signature
462	PS01123	0.0	Thermonuclease family signature 1
463	PS00820	0.0	Glucoamylase active site region signature
464	PS00502	0.04285714285714286	Polygalacturonase active site
465	PS00448	0.38125	Clostridium cellulosome enzymes repeated domain signature
466	PS00773	0.34782608695652173	Chitinases family 19 signature 1
467	PS00128	0.2573099415204678	Alpha-lactalbumin / lysozyme C signature
468	PS00512	0.17647058823529413	Alpha-galactosidase signature
469	PS00927	0.0	Trehalase signature 1
470	PS00719	0.09615384615384616	Glycosyl hydrolases family 2 signature 1

Continued...

#C	#PS	%Helix/Strand	Description
471	PS00608	0.0	Glycosyl hydrolases family 2 acid/base catalyst
472	PS00775	0.0	Glycosyl hydrolases family 3 active site
473	PS01324	0.4109543010752688	Glycosyl hydrolases family 4 signature
474	PS00655	0.0	Glycosyl hydrolases family 6 signature 1
475	PS00592	0.0	Glycosyl hydrolases family 9 active sites signature 1
476	PS00698	0.14035087719298245	Glycosyl hydrolases family 9 active sites signature 2
477	PS00777	0.24999999999999997	Glycosyl hydrolases family 11 active site signature 2
478	PS00953	0.1323529411764706	Glycosyl hydrolases family 25 active sites signature
479	PS00707	0.2903225806451613	Glycosyl hydrolases family 31 signature 2
480	PS00609	0.0	Glycosyl hydrolases family 32 active site
481	PS01140	0.4166666666666667	Glycosyl hydrolases family 45 active site
482	PS60000	0.0	Chitosanases families 46 and 80 active sites signature
483	PS00922	0.27586206896551724	Prokaryotic transglycosylases signature
484	PS00516	0.28	Alkylbase DNA glycosidases alkA family signature
485	PS01242	0.04	Zinc finger FPG-type signature
486	PS00738	0.4000000000000001	S-adenosyl-L-homocysteine hydrolase signature 1
487	PS00739	0.5762527233115469	S-adenosyl-L-homocysteine hydrolase signature 2
488	PS00491	0.0	Aminopeptidase P and proline dipeptidase signature
489	PS00680	0.10526315789473684	Methionine aminopeptidase subfamily 1 signature
490	PS01202	0.5147058823529412	Methionine aminopeptidase subfamily 2 signature
491	PS00869	0.34782608695652173	Renal dipeptidase active site
492	PS00560	0.25555555555555554	Serine carboxypeptidases, histidine active site

Continued...

#C	#PS	%Helix/Strand	Description
493	PS00132	0.5479051383399209	Zinc carboxypeptidases, zinc-binding region 1 signature
494	PS01333	0.1323529411764706	Pyrrolidone-carboxylate peptidase glutamic acid active site
495	PS00672	0.24444444444444446	Serine proteases, V8 family, histidine active site
496	PS00708	0.5612903225806452	Prolyl endopeptidase family serine active site
497	PS00382	0.23626373626373626	Endopeptidase Clp histidine active site
498	PS00640	0.16799999999999998	Eukaryotic thiol (cysteine) proteases asparagine active site
499	PS00140	0.5882352941176471	Ubiquitin carboxyl-terminal hydrolase family 1 cysteine active-site
500	PS00972	0.4479166666666667	Ubiquitin carboxyl-terminal hydrolases family 2 signature 1
501	PS00973	0.21754385964912276	Ubiquitin carboxyl-terminal hydrolases family 2 signature 2
502	PS00141	0.22564102564102573	Eukaryotic and viral aspartyl proteases active site
503	PS00835	0.5882352941176471	Aspartyl proteases, omptin family signature 2
504	PS00143	0.2448671497584541	Insulinase family, zinc-binding region signature
505	PS00388	0.2621788537549407	Proteasome A-type subunits signature
506	PS00854	0.25685890257558797	Proteasome B-type subunits signature
507	PS00760	0.3076923076923077	Signal peptidases I lysine active site
508	PS00571	0.4270833333333333	Amidases signature
509	PS01120	0.0	Urease nickel ligands signature
510	PS00759	0.5476923076923077	ArgE / dapE / ACY1 / CPG2 / yscS family signature 2
511	PS00483	0.8452380952380952	Dihydroorotase signature 2
512	PS00744	0.27384615384615385	Beta-lactamases class B signature 2
513	PS01053	0.11363636363636363	Arginase family signature 3
514	PS00903	0.47430459555649873	Cytidine and deoxycytidylate deaminases zinc-binding region signature
515	PS00387	0.0	Inorganic pyrophosphatase signature
516	PS00150	0.03409090909090909	Acylphosphatase signature 1
517	PS00605	0.5	ATP synthase c subunit signature
518	PS00931	0.0	Cutinase, aspartate and histidine active sites

Continued...

#C	#PS	%Helix/Strand	Description
519	PS00392	0.3181818181818182	DDC / GAD / HDC / TyrDC pyridoxal-phosphate attachment site
520	PS00703	0.0	Orn/Lys/Arg decarboxylases family 1 pyridoxal-P attachment site
521	PS00879	0.09523809523809523	Orn/DAP/Arg decarboxylases family 2 signature 2
522	PS00156	0.42857142857142855	Orotidine 5'-phosphate decarboxylase active site
523	PS00781	0.125	Phosphoenolpyruvate carboxylase active site 1
524	PS00532	0.075	Phosphoenolpyruvate carboxykinase (ATP) signature
525	PS00614	0.2830409356725146	Indole-3-glycerol phosphate synthase signature
526	PS00806	0.4222222222222222	Fructose-bisphosphate aldolase class-II signature 2
527	PS00815	0.35294117647058826	Alpha-isopropylmalate and homocitrate synthases signature 1
528	PS00816	0.42857142857142855	Alpha-isopropylmalate and homocitrate synthases signature 2
529	PS00161	0.0	Isocitrate lyase signature
530	PS00162	0.29411764705882354	Alpha-carbonic anhydrases signature
531	PS00705	0.23809523809523805	Prokaryotic-type carbonic anhydrases signature 2
532	PS00163	0.0	Fumarate lyases signature
533	PS00450	0.1328976034858388	Aconitase family signature 1
534	PS01244	0.0	Aconitase family signature 2
535	PS01028	0.6704301075268817	Dehydroquinase class I active site
536	PS01029	0.0	Dehydroquinase class II signature
537	PS00164	0.28571428571428564	Enolase signature
538	PS00165	0.10535714285714286	Serine/threonine dehydratases pyridoxal-phosphate attachment site
539	PS00166	0.4179894179894179	Enoyl-CoA hydratase/isomerase signature
540	PS00167	0.2857142857142857	Tryptophan synthase alpha chain signature
541	PS01233	0.0	Urocanase signature
542	PS00665	0.0	Dihydrodipicolinate synthetase signature 1
543	PS00666	0.4157006048387097	Dihydrodipicolinate synthetase signature 2

Continued...

#C	#PS	%Helix/Strand	Description
544	PS00901	0.29849624060150376	Cysteine synthase/cystathionine beta-synthase P-phosphate attachment site
545	PS00488	0.5833333333333333	Phenylalanine and histidine ammonia-lyases signature
546	PS00533	0.4117647058823529	Porphobilinogen deaminase cofactor-binding site
547	PS00934	0.45454545454545453	Glyoxalase I signature 1
548	PS00452	0.4166666666666667	Guanylate cyclase signature
549	PS00787	0.2708333333333333	Chorismate synthase signature 1
550	PS00789	0.4529411764705882	Chorismate synthase signature 3
551	PS00987	0.36363636363636365	6-pyruvoyl tetrahydropterin synthase signature 1
552	PS00988	0.0	6-pyruvoyl tetrahydropterin synthase signature 2
553	PS00534	0.2807017543859649	Ferrochelatase signature
554	PS00395	0.0	Alanine racemase pyridoxal-phosphate attachment site
555	PS01326	0.0	Diaminopimelate epimerase signature
556	PS00908	0.7243589743589743	Mandelate racemase / muconate lactonizing enzyme family signature 1
557	PS00909	0.40625	Mandelate racemase / muconate lactonizing enzyme family signature 2
558	PS01085	0.3333333333333333	Ribulose-phosphate 3-epimerase family signature 1
559	PS01086	0.314975845410628	Ribulose-phosphate 3-epimerase family signature 2
560	PS00170	0.007936507936507936	Cyclophilin-type peptidyl-prolyl cis-trans isomerase signature
561	PS01096	0.2424242424242424	PpiC-type peptidyl-prolyl cis-trans isomerase signature
562	PS00965	0.3333333333333333	Phosphomannose isomerase type I signature 1
563	PS00966	0.11538461538461539	Phosphomannose isomerase type I signature 2
564	PS00765	0.0	Phosphoglucose isomerase signature 1
565	PS00174	0.3333333333333333	Phosphoglucose isomerase signature 2

Continued...

#C	#PS	%Helix/Strand	Description
566	PS01161	0.1929824561403509	Glucosamine/galactosamine-6-phosphate isomerases signature
567	PS00544	0.32051282051282054	Methylmalonyl-CoA mutase signature
568	PS01074	0.33333333333333337	Terpene synthases signature
569	PS00176	0.2894736842105263	Eukaryotic DNA topoisomerase I active site
570	PS00396	0.16908212560386474	Prokaryotic DNA topoisomerase I active site
571	PS00178	0.12266666666666667	Aminoacyl-transfer RNA synthetases class-I signature
572	PS00762	0.7758620689655172	WHEP-TRS domain signature
573	PS01216	0.4583333333333333	ATP-citrate lyase / succinyl-CoA ligases family signature 1
574	PS00399	0.0	ATP-citrate lyase / succinyl-CoA ligases family active site
575	PS01217	0.6153846153846154	ATP-citrate lyase / succinyl-CoA ligases family signature 3
576	PS00180	0.08114035087719298	Glutamine synthetase signature 1
577	PS00181	0.27389705882352944	Glutamine synthetase putative ATP-binding region signature
578	PS00182	0.11538461538461539	Glutamine synthetase class-I adenylation site
579	PS00843	0.25	D-alanine–D-alanine ligase signature 1
580	PS00844	0.26826765188834156	D-alanine–D-alanine ligase signature 2
581	PS01011	0.6041666666666666	Folylpolyglutamate synthase signature 1
582	PS00183	0.04588779956427015	Ubiquitin-conjugating enzymes active site
583	PS00565	0.3333333333333333	Argininosuccinate synthase signature 2
584	PS00866	0.13333333333333336	Carbamoyl-phosphate synthase subdomain signature 1
585	PS00333	0.1411111111111111	ATP-dependent DNA ligase signature 2
586	PS01055	0.2	NAD-dependent DNA ligase signature 1
587	PS01056	0.375	NAD-dependent DNA ligase signature 2
588	PS01287	0.0	RNA 3'-terminal phosphate cyclase signature
589	PS01313	0.34375	Lipoate-protein ligase B signature
590	PS01234	0.3111111111111111	Glutamyl-tRNA(Gln) amidotransferase subunit B signature
591	PS00949	0.23076923076923078	Autoinducers synthetase family signature
592	PS00187	0.5833333333333334	Thiamine pyrophosphate enzymes signature
593	PS00188	0.2777777777777778	Biotin-requiring enzymes attachment site
594	PS00189	0.19743589743589748	2-oxo acid dehydrogenases acyltransferase component lipoyl binding site

Continued...

#C	#PS	%Helix/Strand	Description
595	PS00455	0.19444444444444442	Putative AMP-binding domain signature
596	PS01078	0.21428571428571427	Molybdenum cofactor biosynthesis proteins signature 1
597	PS01079	0.23202614379084968	Molybdenum cofactor biosynthesis proteins signature 2
598	PS01235	0.33684210526315783	PdxS/SNZ family signature
599	PS00191	0.0	Cytochrome b5 family, heme-binding domain signature
600	PS00537	0.1	Cytochrome b559 subunits heme-binding site signature
601	PS01000	0.44	Succinate dehydrogenase cytochrome b subunit signature 1
602	PS00195	0.45521390374331544	Glutaredoxin active site
603	PS00196	0.19361044417767112	Type-1 copper (blue) proteins signature
604	PS00814	0.0	Adrenodoxin family, iron-sulfur binding region signature
605	PS00202	0.0	Rubredoxin signature
606	PS00696	0.32098765432098777	Electron transfer flavoprotein alpha-subunit signature
607	PS01065	0.37445887445887444	Electron transfer flavoprotein beta-subunit signature
608	PS00203	0.0	Vertebrate metallothioneins signature
609	PS00204	0.7891156462585034	Ferritin iron-binding regions signature 2
610	PS01344	0.24444444444444446	Frataxin family signature
611	PS00206	0.4318771626297578	Transferrins signature 2
612	PS00207	0.06332138590203107	Transferrins signature 3
613	PS01213	0.36063492063492064	Protozoan/cyanobacterial globins signature
614	PS00550	0.7083333333333333	Hemerythrin family signature
615	PS00210	0.0	Arthropod hemocyanins / insect LSPs signature 2
616	PS01047	0.36688357434186136	Heavy-metal-associated domain
617	PS01037	0.02564102564102564	Bacterial extracellular solute-binding proteins, family 1 signature
618	PS01040	0.2719367588932807	Bacterial extracellular solute-binding proteins, family 5 signature
619	PS00212	0.54	Serum albumin family signature
620	PS00768	0.21428571428571427	Transthyretin signature 1
621	PS00769	0.48717948717948717	Transthyretin signature 2

Continued...

#C	#PS	%Helix/Strand	Description
622	PS00213	0.15	Lipocalin signature
623	PS00214	0.3148148148148148	Cytosolic fatty-acid binding proteins signature
624	PS00400	0.5606060606060606	LBP / BPI / CETP family signature
625	PS01220	0.2173913043478261	Phosphatidylethanolamine-binding protein family signature
626	PS00597	0.06904761904761905	Plant lipid transfer proteins signature
627	PS00897	0.8	LacY family proton/sugar symporters signature 2
628	PS01219	0.5769230769230769	Ammonium transporters signature
629	PS00757	0.3333333333333333	Prokaryotic sulfate-binding proteins signature 2
630	PS00576	0.38235294117647056	General diffusion Gram-negative porins signature
631	PS01068	0.5777777777777777	OmpA-like domain
632	PS01132	0.41538461538461535	Actins and actin-related proteins signature
633	PS00223	0.6042123738481792	Annexins repeated domain signature
634	PS01239	0.4	Dynein light chain type 1 signature
635	PS01134	0.22857142857142856	FtsZ protein signature 1
636	PS00748	0.3333333333333333	F-actin capping protein alpha subunit signature1
637	PS00749	0.0	F-actin capping protein alpha subunit signature 2
638	PS00232	0.27816627816627815	Cadherin domain signature
639	PS00555	0.3461538461538462	Plant viruses icosahedral capsid proteins 'S' region signature
640	PS00236	0.05333333333333333	Neurotransmitter-gated ion-channels signature
641	PS00649	0.032	G-protein coupled receptors family 2 signature 1
642	PS00979	0.47368421052631576	G-protein coupled receptors family 3 signature 1
643	PS00980	0.21739130434782608	G-protein coupled receptors family 3 signature 2
644	PS00238	0.6176470588235294	Visual pigments (opsins) retinal binding site
645	PS00240	0.0	Receptor tyrosine kinase class III signature
646	PS00790	0.3	Receptor tyrosine kinase class V signature 1

Continued...

#C	#PS	%Helix/Strand	Description
647	PS01352	0.3388480171180593	Long hematopoietin receptor, single chain family signature
648	PS01354	0.2888519748984865	Long hematopoietin receptor, soluble alpha chains family signature
649	PS01355	0.34520833333333334	Short hematopoietin receptor family 1 signature
650	PS01356	0.46875	Short hematopoietin receptor family 2 signature
651	PS00652	0.034396866991102025	TNFR/NGFR family cysteine-rich region signature
652	PS00243	0.05357142857142857	Integrins beta chain cysteine-rich domain signature
653	PS00458	0.6944444444444444	Natriuretic peptides receptors signature
654	PS00969	0.7493951612903226	Antenna complexes beta subunits signature
655	PS00430	0.39913127413127414	TonB-dependent receptor proteins signature 1
656	PS01299	0.375	Ephrins signature
657	PS00799	0.0	Granulins signature
658	PS00247	0.11227272727272726	HBGF/FGF family signature
659	PS00619	0.12	PTN/MK heparin-binding protein family signature 1
660	PS00248	0.08571428571428572	Nerve growth factor family signature
661	PS00249	0.0	Platelet-derived growth factor (PDGF) family signature
662	PS00471	0.19620347788710416	Small cytokines (intercrine/chemokine) C-x-C subfamily signature
663	PS00472	0.17722564712988786	Small cytokines (intercrine/chemokine) C-C subfamily signature
664	PS00250	0.06818181818181818	TGF-beta family signature
665	PS00251	0.4159663865546218	TNF family signature
666	PS00252	0.7543859649122807	Interferon alpha, beta and delta family signature
667	PS00253	0.0	Interleukin-1 signature
668	PS00838	0.38873626373626374	Interleukins -4 and -13 signature
669	PS01250	0.3333333333333333	Arthropod CHH/MIH/GIH neurohormones family signature
670	PS00817	0.4074074074074074	Erythropoietin / thrombopoeitin signature

Continued...

#C	#PS	%Helix/Strand	Description
671	PS00260	0.5238095238095238	Glucagon / GIP / secretin / VIP family signature
672	PS00779	0.0	Glycoprotein hormones alpha chain signature 1
673	PS00780	0.0	Glycoprotein hormones alpha chain signature 2
674	PS00689	0.05555555555555555	Glycoprotein hormones beta chain signature 2
675	PS00262	0.13904761904761903	Insulin family signature
676	PS00263	0.0	Natriuretic peptides signature
677	PS00266	0.4411764705882353	Somatotropin, prolactin and related hormones signature 1
678	PS00269	0.07483719983719983	Mammalian defensins signature
679	PS00947	0.0	Cathelicidins signature 2
680	PS00270	0.0	Endothelin family signature
681	PS60011	0.11306614532420985	Plant C6 type antimicrobial peptide (AMP) signature
682	PS00940	0.34268774703557314	Gamma-thionins family signature
683	PS00272	0.20644268012689063	Snake toxins signature
684	PS00459	0.08108108108108109	Myotoxins signature
685	PS01138	0.2664463222933987	Scorpion short toxins signature
686	PS60028	0.06896551724137931	Scorpion calcine family signature
687	PS60015	0.024193548387096774	Mu-agatoxin family signature
688	PS60018	0.0	Delta-atracotoxin (ACTX) family signature
689	PS60017	0.0	Omega-atracotoxin (ACTX) type 2 family signature
690	PS60020	0.1	Janus-faced atracotoxin (J-ACTX) family signature
691	PS60021	0.013888888888888888	Huwentoxin-1 family signature
692	PS60022	0.0	Huwentoxin-2 family signature
693	PS60026	0.2222222222222222	Ergtoxin family signature
694	PS60010	0.05357142857142857	Assassin bug toxin signature
695	PS60014	0.4285714285714286	Alpha-conotoxin family signature
696	PS60004	0.0	Omega-conotoxin family signature
697	PS60005	0.0	Delta-conotoxin family signature
698	PS60019	0.0	I-superfamily conotoxin signature
699	PS60030	0.0625	Bacteriocin class IIa family signature
700	PS00275	0.5705882352941176	Shiga/ricin ribosomal inactivating toxins active site signature
701	PS00481	0.0	Thiol-activated cytolysins signature

Continued...

#C	#PS	%Helix/Strand	Description
702	PS00280	0.17451523545706368	Pancreatic trypsin inhibitor (Kunitz) family signature
703	PS00281	0.0	Bowman-Birk serine protease inhibitors family signature
704	PS00282	0.2412714097496706	Kazal serine protease inhibitors family signature
705	PS00283	0.0	Soybean trypsin inhibitor (Kunitz) protease inhibitors family signature
706	PS00284	0.33057851239669406	Serpins signature
707	PS00285	0.31944444444444436	Potato inhibitor I family signature
708	PS00286	0.0	Squash family of serine protease inhibitors signature
709	PS00999	0.3684210526315789	Streptomyces subtilisin-type inhibitors signature
710	PS00287	0.5119047619047619	Cysteine proteases inhibitors signature
711	PS00426	0.2611440491875275	Cereal trypsin/alpha-amylase inhibitors family signature
712	PS00477	0.14814814814814814	Alpha-2-macroglobulin family thiolester region signature
713	PS00296	0.25	Chaperonins cpn60 signature
714	PS00681	0.15692307692307692	Chaperonins cpn10 signature
715	PS00750	0.35897435897435903	Chaperonins TCP-1 signature 1
716	PS00751	0.29411764705882354	Chaperonins TCP-1 signature 2
717	PS00995	0.0	Chaperonins TCP-1 signature 3
718	PS00329	0.6785714285714286	Heat shock hsp70 proteins family signature 2
719	PS00870	0.5256410256410257	Chaperonins clpA/B signature 1
720	PS00871	0.15789473684210525	Chaperonins clpA/B signature 2
721	PS00636	0.625	Nt-dnaJ domain signature
722	PS01071	0.06818181818181818	grpE protein signature
723	PS01141	0.14814814814814814	Bacterial type II secretion system protein C signature
724	PS00662	0.2	Bacterial type II secretion system protein E signature
725	PS01312	0.1875	SecA family signature
726	PS00755	0.55	Protein secY signature 1
727	PS00756	0.6111111111111112	Protein secY signature 2
728	PS00635	0.4722222222222222	Gram-negative pili assembly chaperone signature

Continued...

#C	#PS	%Helix/Strand	Description
729	PS00300	0.0	SRP54-type proteins GTP-binding domain signature
730	PS00945	0.38636363636363635	Cyclin-dependent kinases regulatory subunits signature 2
731	PS00292	0.8203125	Cyclins signature
732	PS01251	0.33333333333333337	Proliferating cell nuclear antigen signature 1
733	PS00293	0.5263157894736842	Proliferating cell nuclear antigen signature 2
734	PS01080	0.49569377990430613	Apoptosis regulator, Bcl-2 family BH1 motif signature
735	PS01258	0.4772727272727273	Apoptosis regulator, Bcl-2 family BH2 motif signature
736	PS00295	0.18421052631578946	Arrestins signature
737	PS00674	0.40789473684210525	AAA-protein family signature
738	PS00299	0.2867132867132867	Ubiquitin domain signature
739	PS01115	0.0	GTP-binding nuclear protein ran signature
740	PS01270	0.5517241379310345	Band 7 protein family signature
741	PS00741	0.5221153846153845	Dbl homology (DH) domain signature
742	PS00720	0.3617119966266076	Ras Guanine-nucleotide exchange factors domain signature
743	PS00301	0.171875	GTP-binding elongation factors signature
744	PS00825	0.0	Elongation factor 1 beta/beta'/delta chain signature 2
745	PS01126	0.625	Elongation factor Ts signature 1
746	PS01275	0.0	Elongation factor P signature
747	PS01262	0.08695652173913043	Eukaryotic initiation factor 1A signature
748	PS00813	0.15624999999999997	Eukaryotic initiation factor 4E signature
749	PS00302	0.125	Eukaryotic initiation factor 5A hypusine signature
750	PS01176	0.043478260869565216	Initiation factor 2 signature
751	PS00745	0.08823529411764706	Prokaryotic-type class I peptide chain release factors signature
752	PS01319	0.3977272727272727	Ribosome-binding factor A signature
753	PS00803	0.0	Calreticulin family signature 1
754	PS00805	0.0	Calreticulin family repeated motif signature
755	PS00303	0.41761363636363635	S-100/ICaBP type calcium binding protein signature
756	PS00330	0.3947368421052631	Hemolysin-type calcium-binding region signature
757	PS00638	0.28571428571428564	P-II protein C-terminal region signature

Continued...

#C	#PS	%Helix/Strand	Description
758	PS00960	0.0	BTG family signature 1
759	PS01256	0.40151098901098903	Cullin family signature
760	PS01310	0.0	FXYD family signature
761	PS01272	0.3333333333333333	Glucokinase regulatory protein family signature
762	PS00892	0.1929824561403509	HIT domain signature
763	PS00990	0.5526315789473684	Clathrin adaptor complexes medium chain signature 1
764	PS00991	0.0	Clathrin adaptor complexes medium chain signature 2
765	PS00914	0.9591891891891893	Syntaxin / epimorphin family signature
766	PS00307	0.23571428571428568	Legume lectins beta-chain signature
767	PS00985	0.25333333333333335	Spermadhesins family signature 1
768	PS00986	0.21212121212121213	Spermadhesins family signature 2
769	PS00621	0.08333333333333333	Tissue factor signature
770	PS01002	0.0	Translationally controlled tumor protein signature 1
771	PS01003	0.17391304347826086	Translationally controlled tumor protein signature 2
772	PS01200	0.0	Tub family signature 1
773	PS00305	0.0	11-S plant seed storage proteins signature
774	PS60008	0.0	Cyclotides bracelet subfamily signature
775	PS00725	0.0	Germin family signature
776	PS00451	0.4787878787878788	Pathogenesis-related proteins Bet v I family signature
777	PS00316	0.03529411764705882	Thaumatin family signature
778	PS01281	0.625	Glucose inhibited division protein A family signature 2
779	PS01228	0.3333333333333333	Hypothetical cof family signature 1
780	PS01229	0.30434782608695654	Hypothetical cof family signature 2
781	PS01211	0.4	Uncharacterized protein family UPF0001 signature
782	PS01246	0.45714285714285713	Uncharacterized protein family UPF0003 signature
783	PS01227	0.40476190476190477	Uncharacterized protein family UPF0012 signature
784	PS01267	0.0	Uncharacterized protein family UPF0023 signature

Continued...

#C	#PS	%Helix/Strand	Description
785	PS01269	0.30000000000000004	Uncharacterized protein family UPF0025 signature
786	PS00910	0.4666666666666667	Uncharacterized protein family UPF0029 signature
787	PS01148	0.25	Uncharacterized protein family UPF0033 signature
788	PS01318	0.13636363636363635	Uncharacterized protein family UPF0066 signature
789	PS01320	0.08333333333333333	Uncharacterized protein family UPF0067 signature
790	PS01094	0.047368421052631574	Uncharacterized protein family UPF0076 signature
791	PS01152	0.2222222222222222	Hypothetical hesB/yadR/yfhF family signature