ALGORITHMIC DEVELOPMENTS FOR SEQUENCE ANALYSIS, STRUCTURE MODELING AND FUNCTIONAL PREDICTION OF PROTEINS

APPROVED BY SUPERVISORY COMMITTEE

Nick V. Grishin, Ph.D.; Advisor

Alexander Pertsemlidis, Ph.D.; Committee Chair

Stephen R. Sprang, Ph.D.

Zbyszek Otwinowski, Ph.D.

I would like to thank the members of my Graduate Committee and the Graduate school.

To my daughter Erica, my husband Yi, my mother, father and brother.

ALGORITHMIC DEVELOPMENTS FOR SEQUENCE ANALYSIS, STRUCTURE MODELING AND FUNCTIONAL PREDICTION OF PROTEINS

by

YUAN QI

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences
The University of Texas Southwestern Medical Center at Dallas
In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
The University of Texas Southwestern Medical Center at Dallas
Dallas, Texas
November 2006

Copyright

by

Yuan Qi, 2006

All Rights Reserved

ALGORITHMIC DEVELOPMENTS FOR SEQUENCE ANALYSIS, STRUCTURE MODELING AND FUNCTIONAL PREDICTION OF PROTEINS

Publication No. ________________

Yuan Qi, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2006

Supervising Professor: Nick V. Grishin, Ph.D.

Sequence, structure and function, being the three most important properties of proteins, are interrelated through homology relationships. In this post-genome era, we are equipped with abundant sequence information. Homology inference is thus of great practical importance because of its ability to make structural and functional predictions through sequence analysis. In an effort to explore and utilize the protein sequence-structure-function relationships, with homology detection and utilization as the central scheme, this work concentrates on algorithmic development of methods and systems for sequence similarity search, structure modeling and functional prediction purposes, as well as performs structure prediction and classification for specific protein families.

Three algorithmic developments are described in this dissertation. First, to facilitate identification of structurally or functionally important interactions between positions in a protein family, a program has been developed to perform positional correlation analysis of multiple sequence alignments using different methods. The program has been shown to be useful to identify functionally important position pairs or networks of correlated positions.

Second, to further increase the sensitivity of sequence similarity search methods in terms of homology detection and structure modeling ability, a method has been developed by incorporating predicted secondary structure information with sequence profiles. Evaluation on PFAM-based system shows that this method provides improved structure template detection ability and generates alignment of better quality.

Third, in order to systematically assess the structure modeling abilities of different sequence similarity search programs, a comprehensive evaluation system has been developed. This large-scale automatic evaluation system assesses the fold recognition ability and alignment quality of different programs from global and local perspectives using both reference-dependent and reference-independent approaches, which provides an instrument to understand the progress and limitations of the field.

Two structure prediction and classification projects using manual analysis and existing tools are also described in this dissertation. First, the structure of C-terminal domain of Gyrase A is predicted through inferred homology relationship with regulator of chromosome condensation (RCC1). This prediction has been validated by experimental data. Second, a hierarchical structure classification of thioredoxin-like fold proteins has been carried out, which promotes understanding of fold definitions and sequence-structure-function relationships.

TABLE OF CONTENTS

Committee Signatures i

Dedication ii

Title Page iii

Copyright iv

Abstract v

Table of Contents vii

Prior Publications xiv

List of Figures xv

List of Tables xviii

CHAPTER 1: General Introduction 1

1.1 Homology Interrelates Protein Sequence, Structure And Function 1

1.1.1 Homology Detection And Sequence Analysis 1

1.1.2 Structure Modeling 2

1.1.3 Functional Prediction 4

1.1.4 Structure Classification 5

1.2 Overview of dissertation work 6

CHAPTER 2: Structure Prediction of C-Terminal Domain of Gyrase A 8

2.1 Introduction 8

2.1.1 Background 8

2.1.2 Objective 9

2.2 Materials and Methods 10

2.2.1 Sequence Similarity Searches ... 10

2.2.2 Multiple Sequence Alignment And Hydrophobicity Analysis ... 10

2.2.3 Secondary Structure Prediction And Threading ... 11

2.3 Results ... 12

2.3.1 PSI-BLAST Searches ... 12

2.3.2 Secondary Structure Predictions And Fold Recognition ... 13

2.3.3 Multiple Sequence Alignment ... 14

2.4 Discussion ... 15

2.4.1 Validity Of The Fold Prediction ... 15

2.4.2 Structural Differences Between Gyra/Parc And RCC1 ... 16

2.4.3 Functional Implications ... 17

2.4.4 Prediction Confirmation ... 18

2.5 Conclusions ... 18

CHAPTER 3: Structural Classification of Thioredoxin-like Fold Proteins ... 27

3.1 Introduction ... 27

3.1.1 Background ... 27

3.1.2 Objective ... 28

3.2 Materials and Methods ... 29

3.2.1 Structural Motif Search For Thioredoxin-Like Protein Domains ... 29

3.2.2 Sequence-Based Classification Of The Thioredoxin-Like Protein Domains ... 30

3.2.3 Structure And Function Based Classifications ... 31

3.3 Overall Fold Description ... 31

3.3.1 Thioredoxin-like fold 31

3.3.2 Circular Permutations 33

3.4 Description Of Thioredoxin-Like Fold Families 34

3.4.1 Thioredoxin family 35

3.4.2 RTPC small domain family 36

3.4.3 Ribosomal protein L30e family 36

3.4.4 Tubulin C-terminal domain family 37

3.4.5 Bacillus chorismate mutase (BCM) Family 37

3.4.6 MECP synthase family 38

3.4.7 PurM N-terminal domain family 38

3.4.8 Cytidine deaminase family 39

3.4.9 AICAR Tfase domain of bifunctional purine biosynthesis enzyme ATIC family 40

3.4.10 Phospholipase D family 40

3.4.11 gp5 domain A family 41

3.5 Discussion 41

3.5.1 Analysis of active site locations 41

3.5.2 Comparison to other structure classifications 43

3.5.3 Structural analogs 45

3.6 Conclusions 47

CHAPTER 4: PCOAT: A Tool For Protein Positional Correlation Analysis 55

4.1 Introduction 55

4.1.1 Background 55

4.1.2 Objective 55

4.2 Algorithms 56

4.2.1 Estimation of effective counts 56

4.2.2 Identification of statistically significantly correlated position pairs and amino acid pairs 57

4.2.3 Identification of highly correlated positions 58

4.2.4 Identification of networks of correlated positions 59

4.3 Applications and Comparison to Other Methods 59

4.4 Program availability 60

4.5 Conclusions 60

CHAPTER 5: Combining Sequence Profile With Predicted Secondary Structure For Structure Modeling And Homology Detection 64

5.1 Introduction 64

5.1.1 Background 64

5.1.2 Objective 65

5.2 Algorithm Development 65

5.2.1 Construction Of Substitution Matrices 66

5.2.2 Effective Count Calculation For Multiple Sequence Alignment 68

5.2.3 Predicted Secondary Structure Information 69

5.2.4 Estimation Of Target Frequencies 70

5.2.5 Scoring System for Amino Acid And Secondary Structure Profiles 71

5.2.6 Estimation Of Statistical Significance Of Optimal Alignments 74

5.3 Results and Discussion of the Substitution Matrices 76

5.4 Results and Discussion of Pfam-based Performance Evaluation 79

5.4.1 Evaluation of Alignment Quality 79

5.4.2 Evaluation of Homology Detection Ability 82

5.5 Result and Discussion of The 60x60 Scoring System Approach 84

5.5.1 Study Of The Substitution Matrix Effects 85

5.5.2 Study Of The Scoring Function Effects 86

5.5.3 Study Of The Effective Count Effects 88

5.6 Applications 89

5.7 Conclusions 91

CHAPTER 6: A Comprehensive System to Evaluate Structure Modeling Ability of Sequence Similarity Search Methods 113

6.1 Introduction 113

6.1.1 Background 113

6.1.2 Objective 114

6.2 Evaluation Algorithm Development 114

6.2.1 Selection of Representative Dataset 114

6.2.2 Reference-Dependent Evaluation Of Structure Template Quality 117

6.2.3 Reference-Dependent Evaluation Of Alignment Quality 119

6.2.4 Summary Of Reference-Dependent Evaluation Criteria 120

6.2.5 Reference-Independent Global Mode Evaluation 121

6.2.6 Reference-Independent Local Mode Evaluation 121

6.2.7 Summary Of Reference-Independent Evaluation Criteria 123
6.3 Results and Discussions of Evaluations of Sequence Similarity Search Programs.... 123
6.3.1 Evaluations On A Small Testing Set 124
6.3.2 Evaluations On The Entire Representative Set... 126
6.4 Results And Discussions Of The Evaluation System 128
6.4.1 Representative Dataset... 130
6.4.2 Global/local mode 130
6.4.3 Reference-dependent evaluation 131
6.4.4 Reference-independent evaluation 131
6.4.5 Why GDT_TS and LiveBench 3dscore Give Similar Results 132
6.4.6 The Model Of Length Dependency Of Local GDT_TS Scores 132
6.5 Conclusions.. 133
CHAPTER 7: Summary and Future Directions.. 150
7.1 Concluding Remarks: Structure Prediction Of Gyrase A C-Terminal Domain 150
7.1.1 Project Summary.. 150
7.1.2 Applications and Utility 151
7.2 Concluding Remarks: Structure Classification Of Thioredoxin-Like Fold Proteins.. 151
7.2.1 Project Summary.. 151
7.2.2 Applications and Utility 151
7.3 Concluding Remarks: Positional Correlation Analysis Algorithm............................ 152
7.3.1 Project Summary.. 152
7.3.2 Applications and Utility 152

7.4 Concluding Remarks: Sequence Similarity Search Method 153
    7.4.1 Project Summary 153
    7.4.2 Applications and Utility 153
7.5 Concluding Remarks: Evaluation System For Sequence Similarity Search Methods 154
    7.5.1 Project Summary 154
    7.5.2 Applications and Utility 154
APPENDIX A Formulae 155
BIBLIOGRAPHY 156
VITAE 164

PRIOR PUBLICATIONS

Qi, Y., and Grishin, N. V. (2005). Structural classification of thioredoxin-like fold proteins. Proteins *58*, 376-388.

Qi, Y., and Grishin, N. V. (2004). PCOAT: positional correlation analysis using multiple methods. Bioinformatics *20*, 3697-3699.

Cheek, S., Qi, Y., Krishna, S. S., Kinch, L., and Grishin, N. V. (2004). SCOPmap: Automated assignment of protein structures to evolutionary superfamilies. BMC Bioinformatics *5*, 197.

Kinch, L. N., Qi, Y., Hubbard, T. J., and Grishin, N. V. (2003). CASP5 target classification. Proteins *53 Suppl 6*, 340-351.

Kinch, L. N., Wrabl, J. O., Krishna, S. S., Majumdar, I., Sadreyev, R. I., Qi, Y., Pei, J., Cheng, H., and Grishin, N. V. (2003). CASP5 assessment of fold recognition target predictions. Proteins *53 Suppl 6*, 395-409.

Qi, Y., Pei, J., and Grishin, N. V. (2002). C-terminal domain of gyrase A is predicted to have a beta-propeller structure. Proteins *47*, 258-264.

Abbott, J. J., Pei, J., Ford, J. L., Qi, Y., Grishin, V. N., Pitcher, L. A., Phillips, M. A., and Grishin, N. V. (2001). Structure prediction and active site analysis of the metal binding determinants in gamma -glutamylcysteine synthetase. J Biol Chem *276*, 42099-42107.

Patra, G., Williams, L. E., Qi, Y., Rose, S., Redkar, R., and Delvecchio, V. G. (2002). Rapid Genotyping of Bacillus anthracis Strains by Real-Time Polymerase Chain Reaction. Ann NY Acad Sci *969*, 106-111.

Qi, Y., Patra, G., Liang, X., Williams, L. E., Rose, S., Redkar, R. J., and DelVecchio, V. G. (2001). Utilization of the rpoB gene as a specific chromosomal marker for real-time PCR detection of *Bacillus anthracis*. Appl Environ Microbiol *67*, 3720-3727.

LIST OF FIGURES

Figure 2.1 Domain compositions of gyrase and topoisomerase IV 20
Figure 2.2 Multiple sequence alignment of C-terminal of GyrA/RCC1 domain 21
Figure 2.3 Structure Diagrams and Conservation Mapping of RCC1 and GyrA/ParC-CTD. 23
Figure 2.4 Experimentally Determined Structure Diagram of GyrA-CTD 24
Figure 3.1 Thioredoxin-like Fold and Its Observed Circular Permutations 48
Figure 3.2 Structure-based multiple sequence alignment of representative thioredoxin-like protein domains.. 51
Figure 3.3 Active site locations 52
Figure 3.4 Structure Analogs 53
Figure 4.1 Structure Diagram of C2H2 Zinc Finger Showing Correlated Pairs..................... 61
Figure 4.2 Structure Diagram of ACT Domain Showing Correlated Network 62
Figure 5.1 My Amino Acid Substitution Matrix AA80 Matches BLOSUM62 92
Figure 5.2 Distribution of Optimal Random Scores Fitted to EVD 93
Figure 5.3 Dependency of Lambda and *K* on Alignment Length and Effective Sequence Number 94
Figure 5.4 Evaluation Results of Alignment Quality.. 95
Figure 5.5 Example of Alignments Generated by Our Method and COMPASS 96
Figure 5.6 Better Accuracy Than COMPASS when Coverage Equals 1 97
Figure 5.7 Evaluation Results of Homology Detection Ability... 98
Figure 5.8 Example of Improved Homology Detection Ability and Alignment Quality 99
Figure 5.9 Example Of New Similarity Detected Between SCOP Superfamilies................ 100
Figure 5.10 Example Of New Similarity Detected Between SCOP Folds 101
Figure 5.11 Comparison of BLOSUM62 and Our Amino Acid Substitution Matrix 102
Figure 5.12 The Three Substitution Matrices We Calculated... 103
Figure 5.13 Secondary Structure Propensities Extracted from 60x60 Substitution Matrix.. 104
Figure 5.14 Homology Detection Ability of 60x60 Approach Fails 105
Figure 5.15 ROC Curve Analysis and Family-based Paired T-Test Result of Eight Pfam Families 106

Figure 5.16 Distributions of Amino Acids and Predicted Secondary Structure Elements in Blocks+ Database 107
Figure 5.17 Comparison of the 60x60 Substitution Matrix and the Independent Matrix..... 108
Figure 5.18 Comparison of Homology Detection Ability Effects of 60x60 Substitution Matrix and Independent Matrix 109
Figure 5.19 Comparison of Homology Detection Ability Effects of 60x60 Scoring System and 20x20+3x3 Scoring System with Equal Weights 1 110
Figure 5.20 Comparison of Homology Detection Ability Effects of Weighted 60x60 Scoring System and the 20x20+3x3 Scoring System.. 111
Figure 5.21 Effects of the Number of Effective Count.. 112
Figure 6.1 Graphical Representation of Domain Relationship within A Superfamily......... 134
Figure 6.2 Representatives Selected Using Different Structure Alignment Programs and Different Identity Measures 135
Figure 6.3 SVM Score Cutoff Selections for Overall Structure Template Quality............. 136
Figure 6.4 Cutoff Selections for Alignment Quality 137
Figure 6.5 Flowchart of Reference-dependent Evaluation System 138
Figure 6.6 Length-dependency of GDT_TS 139
Figure 6.7 Local GDT_TS Cutoff Method Selection 140
Figure 6.8 Flowchart of Reference-independent Evaluation System 141
Figure 6.9 Distributions of Representative Domains.. 142
Figure 6.10 Reference-dependent Structure Template Quality Evaluation of Various Sequence Alignment Programs on the Testing Set.. 143
Figure 6.11 Reference-dependent Alignment Quality Evaluation of Various Sequence Alignment Programs on the Testing Set.. 144
Figure 6.12 Reference-independent Global Mode Evaluation of Various Sequence Alignment Programs on the Testing Set 145
Figure 6.13 Reference-independent Local Mode Evaluation of Various Sequence Alignment Programs on the Testing Set 146

Figure 6.14 Evaluation Results of Various Sequence Alignment Programs on the Entire Representative Dataset 147
Figure 6.15 Comparison of GDT_TS and LiveBench 3dscore Functions 148

LIST OF TABLES

Table 2.1 Average hydrophobicity of beta-strands in GyrA/ParC and RCC1 25
Table 2.2 Comparison Between Prediction and Experimental Data....................................... 26
Table 3.1 Summary of Thioredoxin-like Domain Families 54
Table 4.1 Comparison of Analysis Results Using PCOAT, DEPENDENCY and CRASP ... 63
Table 5.1 Relative Entropy and Expected Scores of Substitution Matrices 76
Table 5.2 Number of Similar Pfam family-pairs Predicted 90
Table 6.1 SVM Score Cutoffs 149

CHAPTER 1:
General Introduction

1.1 HOMOLOGY INTERRELATES PROTEIN SEQUENCE, STRUCTURE AND FUNCTION

Sequence, structure and function are three major properties of proteins. All studies about proteins are essentially developed around these three aspects. From the evolutionary point of view, homology plays a central role and interrelates these three properties of proteins. Through inferred homology relationships, protein structure and function can be predicted from protein sequence, and sequence similarities can be supported or verified from structural or functional features. We discuss the details of their relationships in the following sections.

1.1.1 Homology Detection And Sequence Analysis

Homologous proteins are proteins that have evolved from the same ancestor. Homologous proteins typically possess conserved sequence motifs and structural features, the same structure folds, and similar functional sites and general biochemical functions. Strong sequence similarity alone or combined sequence-structure or sequence-function similarities are often used to establish homology relationships between proteins. Distinctive structure features, similar structural folds, conserved sequence motifs and functional sites are often used to further support or verify the inference of homology.

Many sequence similarity search methods are available to automatically detect homologs from statistically significant sequence information, the most popular one being PSI-BLAST. Protein sequence similarity search methods have advanced greatly over the last one and a half decades. Sequence similarity search methods have been developed from single sequence vs. sequence methods such as BLAST (Altschul, Gish et al. 1990), to sequence vs.

profile methods such as PSI-BLAST (Altschul, Madden et al. 1997) and RPS-BLAST (Marchler-Bauer, Panchenko et al. 2002), to profile vs. profile methods such as BASIC (Rychlewski, Zhang et al. 1998), Prof_sim (Yona and Levitt 2002) and COMPASS (Sadreyev and Grishin 2003), and the ability to detect distant homology has increased greatly. PSI-BLAST (a sequence-profile method) can be used to reliably detect homologs at >30% sequence identity level. Profile-profile based methods can be used to detect homologs at ~20-30% identity level (Sadreyev and Grishin 2003). However, since protein sequences evolve very fast, detecting more remote sequence similarities (< 20% identity) is difficult. Therefore, it is necessary to develop more powerful distant similarity detection method.

Homologous proteins can be grouped together to form protein families. The direct advantages of grouping are (a) the ease of finding annotated sequence neighbors, which is useful in single unknown sequence analysis, and (b) the ability to study the protein family as a whole, which enables the identification of conserved sequence motifs or structure features. Many protein sequence family databases exist (Henikoff, Henikoff et al. 1999; Attwood, Bradley et al. 2003; Marchler-Bauer, Anderson et al. 2005; Hulo, Bairoch et al. 2006; Letunic, Copley et al. 2006) with Pfam as the major one. Pfam (Protein domain families) (Bateman, Coin et al. 2004) is a database of multiple sequence alignments of protein families or conserved protein regions. The multiple sequence alignments are built from seed alignments followed by profile hidden Markov models. Pfam is the largest available source of accurate semiautomatic multiple sequence alignments (Sadreyev and Grishin 2003).

1.1.2 Structure Modeling

Protein structures usually evolve slower than their sequences. Even if their sequences have evolved beyond recognition, homologous proteins could still share similar structure folds. Therefore, the unknown tertiary structure of a protein can be modeled based on the known structures of their homologous proteins.

In practice, when we have a protein sequence of unknown structure, the first step in homology modeling is to search for similar protein sequences with known structures. Once found, we need to infer if the structure-known protein and the structure-unknown protein are homologous or not from the degree of sequence similarities between them. If they are homologous, we will be able to model the structure of the unknown protein based on the known one according to the sequence alignment between them.

There are a few commonly used terms in the field of homology modeling. “Query” or “target” refers to the protein of unknown structure and which starts the sequence similarity search process in order to find similar protein sequences with known structures. “Hit” or “template” refers to the protein with known structure and which is used as a structure model for the query. In some cases, for instance when testing a sequence similarity search program, the query can have known structure. But in all cases, it is the one whose structure needs to be modeled based on that of the template.

The sequence similarity search methods that are discussed in the previous section can be used for homology modeling purposes. Homology modeling can be divided into two categories. If the sequence similarity between the target and the template is strong (>30% identity), the template can be readily found by BLAST or PSI-BLAST. Homology modeling at this sequence similarity level is also called comparative modeling (Tress, Tai et al. 2005). If the template can only be found by profile-profile based or more powerful searches, homology modeling is categorized as fold recognition (Tress, Tai et al. 2005).

Despite the presence of other types of structure prediction methods, such as *ab initio* methods, homology modeling methods for protein structure prediction are of great practical importance (Lattman 2005). In addition to model overall structural folds for unknown sequences, homology modeling can also be used to model active sites or interaction surfaces of proteins with other molecules, and thus has great potential in drug design. Since many sequence similarity search methods exists, it is important to have an evaluation system to assess their performance in terms of homology modeling abilities.

1.1.3 Functional Prediction

Homologous proteins typically share similar functional sites and general biochemical functions. Functional sites or active sites include catalytic sites, substrate-binding sites, or “hot spots” on protein-protein interaction surfaces. Functional sites are identifiable through the multiple sequence alignment of a given protein family. These sites form conserved columns in the multiple sequence alignment since they have evolved under evolutionary selection pressure. When proteins in a given family possess different substrate specificities, they may have different amino acid types conserved at the same sites or have a shift in the position of the conserved sites.

Homologs can be separated into orthologs and paralogs. Orthologs are homologs resulting from speciation events; while paralogs are homologs resulting from gene duplication events (Fitch 2000). Orthologs are believed to have the same function and often the same specificity since they have been under similar evolutionary pressure. On the other hand, paralogs are believed to have diverged to evolve new specificities or even new functions since they have experienced weaker evolutionary pressure after duplication (Mirny and Gelfand 2002). Therefore, identification of orthologs is crucial for reliable protein functional prediction. The databases Clusters of Orthologous Groups of proteins (COG) (Tatusov, Fedorova et al. 2003) and Eukaryotic Orthologous Groups (KOG) (Tatusov, Fedorova et al. 2003) fulfill this intention. COG is constructed from the complete genome sequences of prokaryotes and unicellular eukaryotes, and KOG is constructed from complete genome sequences of eukaryotes. Since orthologous proteins typically have the same function, COG and KOG allow functional information transfer from one member to an entire group. The approach of COG and KOG should facilitate functional annotation of genomes.

Mirny and Gelfand use the concepts of ortholog and paralog to identity specificity-determining residues in bacterial transcription factors (Mirny and Gelfand 2002). They group the orthologous transcription factors together, which are assumed to have the same specificity, and thus transcription factors between groups are considered paralogs. The specificity-determining residues are found by comparing the sequences in different groups.

Mirny and Gelfand's method is similar to the approach of positional correlation analysis, which has been shown to identify specificity-determinant residues (Crowder, Holton et al. 2001).

1.1.4 Structure Classification

More than 39,000 experimentally determined protein structures have been deposited in the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000) and the acceleration in the growth of structure data is anticipated as high-throughput structural genomics continues (Westbrook, Feng et al. 2003). To systematize this large amount of data for better understanding of protein evolution and sequence-structure-function relationships, protein structure classification is necessary. In a protein structure classification, fold group and evolutionary family are the two major levels. At the fold level, protein domains are grouped based on the connectivity and mutual orientation of their core secondary structure elements. Within each fold group, proteins are further divided into evolutionary families based on inferred homology relationships. The same structural folds possessed by non-homologous proteins are considered to be the result of convergent evolution. Protein structure classification is hierarchical in nature, for homologous proteins typically have the same structure fold. The geometry of protein structures usually reflects certain constraints from sequence and function. Thus grouping proteins by folds will aid in understanding of the physico-chemical principles behind protein structures, which in turn could help to address problems such as protein folding and structure-functional prediction. Furthermore, although a few exceptional examples exist where homologous proteins have evolved different folds (Murzin 1998; Grishin 2001), protein structures generally evolve slower than their sequences. Consequently, grouping protein domains by folds could also help in understanding protein evolution and will facilitate homology inference. Therefore, hierarchical protein structure classifications usually take into consideration both structural and evolutionary criteria.

Several hierarchical protein structure classifications exist, with the major ones being SCOP (Murzin, Brenner et al. 1995), CATH (Orengo, Michie et al. 1997) and Dali Domain Dictionary (DaliDD) (Holm and Sander 1996; Holm and Sander 1998; Dietmann and Holm 2001). SCOP (Structural Classification Of Proteins) is constructed by combining expert curation and automatic sequence comparison methods. There are four major levels in the SCOP hierarchy. Starting from the lowest level, a family contains proteins that are close homologs. A superfamily contains families that are remotely homologous to each other. A fold contains superfamilies that share the same structure fold, i.e. the same core secondary structures with the same connectivity and mutual orientation. A class contains folds of the same secondary structure composition (e.g. all alpha or all beta). CATH and DaliDD are constructed using fully automatic methods and have similar hierarchical levels as SCOP.

FSSP (Families of Structurally Similar Proteins) (Holm and Sander 1996) is a non-hierarchical structure classification database, which provides structurally aligned families of proteins based on significant structural similarity. This database is constructed and updated by all-against-all structure comparisons of protein structures in the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000) using the DALI structure comparison program (Holm and Sander 1995).

1.2 OVERVIEW OF DISSERTATION WORK

This dissertation work attempts to explore all aspects of the homology interrelated protein sequence-structure-function relationships discussed above. With identification and utilization of homology relationships as the central scheme, this dissertation work includes structure modeling (Chapter 2), structure classification (Chapter 3), algorithm developments of positional correlation-based functional prediction method (Chapter 4), sequence similarity search method (Chapter 5), and evaluation system to assess the homology modeling performance of sequence similarity search programs (Chapter 6). These projects are described in detail in the following chapters.

In addition to these, project that is not described in this dissertation but also demanded a significant amount of time and work is SCOPlink. SCOPlink (unpublished) is an extension and application of the SCOPmap project (Cheek, Qi et al. 2004). In the SCOPlink project, potential homology relationships between SCOP superfamilies (Murzin, Brenner et al. 1995) are identified by comparing SCOP superfamilies (version 1.65 and 1.69) to each other in an all-against-all fashion. The resulting data, including sequence and structural alignments, are transformed automatically into user-friendly formats and are presented in a web interface for easy browsing and manipulation. Initial curation of the data revealed numerous interesting examples of previously unrecognized homology relationships and networks of related SCOP superfamilies.

CHAPTER 2:
Structure Prediction of C-Terminal Domain of Gyrase A

2.1 INTRODUCTION

2.1.1 Background

Topoisomerases are ubiquitous enzymes that catalyze cleavage and religation of DNA molecules allowing for the changes in DNA topological states (Caron and Wang 1994; Wang 1996). Topoisomerases are involved in crucial cellular processes such as replication, transcription, and recombination, and thus have pharmaceutical importance (Maxwell 1992; Hiasa, Yousef et al. 1996). Topoisomerases of type I and type II cleave one and two DNA strands, respectively. Type II enzymes require ATP for their activity and possess an ATPase domain or subunit. Most bacteria have two homologous type II enzymes: DNA gyrase (topoisomerase II, Gyr) and topoisomerase IV (Par). Each enzyme is composed of two subunits (Figure 2.1). GyrA is involved in breakage and reunion of DNA and GyrB functions as an ATPase. Equivalent subunits in topoisomerase IV, ParC and ParE, share about 35% identity with GyrA and GyrB. Despite pronounced sequence similarity, gyrase and topo IV possess distinct cellular functions (Zechiedrich, Khodursky et al. 2000; Deibler, Rahmati et al. 2001). Gyrase introduces negative supercoils into DNA. Topo IV relaxes negative and positive DNA supercoils (Deibler, Rahmati et al. 2001).

The reaction mechanism of type II topoisomerases is relatively well understood and crystal structures for most of their domains are available. GyrB can be divided into two fragments (Figure 2.1a). The 43kDa N-terminal portion of the *E. coli* enzyme with known structure is composed of an ATPase domain related to MutL/Hsp90/histidine kinase and a ribosomal protein S5-like domain (Murzin, Brenner et al. 1995; Lo Conte, Ailey et al. 2000; Deibler, Rahmati et al. 2001). The 47kDa C-terminal portion consists of a toprim Rossmann-like domain interrupted by an insertion and is homologous to the N-terminal segment of the yeast topoisomerase II with available structure (Aravind, Leipe et al. 1998; Berger, Fass et al. 1998). Domain architecture of ParE is similar except that the insertion in the toprim domain is shorter (Figure 2.1b).

GyrA is also composed of two fragments (Figure 2.1a). The structure of the 59K N-terminal fragment for *E. coli* enzyme has been determined and the position of the catalytic tyrosine has been localized (Morais Cabral, Jackson et al. 1997). The C-terminal 38K fragment of GyrA still remains the largest piece of the topoisomerase sequence without structural information. It has been shown that the C-terminal fragment can be expressed separately. It lacks catalytic activity, but can complement the N-terminal fragment upon mixing, which increases its supercoiling activity(Reece and Maxwell 1991). The C-terminal fragment acts as a non-specific DNA-binding protein and is probably involved in stabilization of the DNA-topoisomerase complex (Reece and Maxwell 1991). Without spatial structure information, this fragment remains poorly understood.

Regulator of chromosome condensation (RCC1) is the guanine-nucleotide-exchange factor for the nuclear G protein, Ran, which controls nucleocytoplasmic transport, mitotic spindle formation, and nuclear envelope assembly (Nemergut 2001). These functions depend on the association of RCC1 with DNA. Mutations in the yeast RCC1 gene affect pre-messenger RNA processing and transport, mating, initiation of mitosis and chromatin decondensation. The crystal structure of RCC1 revealed that the molecule folds as a 7-bladed β-propeller, composed of seven four–stranded β-sheets (blades) arranged in a circular array (Renault, Nassar et al. 1998). The β-propeller proteins vary in the number of blades (from 4 to 8), share limited sequence similarity despite pronounced structural resemblance, and display extreme functional diversity (Paoli 2001).

2.1.2 Objective

In order to help fully understand its biological activities and functions, we decide to prediction the spatial structure of the C-terminal domain of GyrA. Using consensus of probabilistic sequence comparison methods combined with hydrophobicity analysis, we detect sequence similarity between the C-terminal domain of bacterial gyrase A and regulator of chromosome condensation (RCC1) (Renault, Nassar et al. 1998) and infer homology between them. We predict that GyrA/ParC C-terminal domain folds as a 6-bladed β-propeller. Functional implications of this homology prediction are discussed.

2.2 MATERIALS AND METHODS

2.2.1 Sequence Similarity Searches

The PSI-BLAST program was used to search for homologues of the gyrase C-terminal fragment (Altschul, Madden et al. 1997). Residues 510-836 of *Mycoplasma genitalium* GyrA (gi|1346233) were selected as a query to search against the non-redundant (nr) database at NCBI (February 2001, 616,977 sequences, 195,057,269 total letters). The E-value threshold was set to 0.02. All other parameters were defaults (Altschul, Madden et al. 1997). PSI-BLAST was iterated until convergence. Found homologues were grouped by single linkage clustering (BLAST score threshold of 1 bit per site corresponding to about 50% identity) as implemented in the SEALS package (Walker and Koonin 1997), and the representative sequences were used as new queries for subsequent PSI-BLAST iterations.

2.2.2 Multiple Sequence Alignment And Hydrophobicity Analysis

Multiple sequence alignments were constructed using the T-COFFEE program (Notredame, Higgins et al. 2000) and adjusted manually based on the secondary structure predictions (discussed below) and the conserved residue patterns. Alignments for topo II sequences and RCC1 sequences were made separately and then merged based on the PSI-BLAST local alignments and hydrophobicity profiles. Propeller blades corresponding to sequence repeats were aligned to each other. The average hydrophobicity of residues in each of the four β-strands of the blades was calculated separately for topo II and RCC1 alignments using the scale from the mean values of 127 different hydrophobicity scales (Palliser and Parry 2001).

2.2.3 Secondary Structure Prediction And Threading

Five representative (most diverse) topoisomerase C-terminal domain sequences (gi|68494, residues 537–875; gi|1346229, residues 538–922; gi|1346233, residues 514–836; gi|1835202, residues 528–907; gi|6655026, residues 517-755) were submitted to the JPRED2 consensus secondary structure prediction server (http://jura.ebi.ac.uk:8888/) (Cuff and Barton 2000), which returns the consensus of prediction results for six different secondary structure prediction methods, including PHD, NNSSP, DSC, PREDATOR, MULPRED and ZPRED. These five sequences were also submitted to another secondary structure prediction server, SAM-T99 (http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html) (Karplus, Barrett et al. 1998). JPRED2 secondary structure predictions were also carried out for 3 RCC1 sequences (gi|12325184, gi|132174, gi|7493765).

Seven fold recognition (threading) methods were applied to five representatives of the gyrase C-terminal domain (gi|68494, gi|1346229, gi|121882, gi|1346235, gi|729651). The following methods were explored: (1) the hybrid fold recognition method of Fischer at the BioInBgu server (http://www.cs.bgu.ac.il/~bioinbgu/) (Fischer 2000); (2) a method that combines multiple sequence profiles and knowledge of protein structures to provide enhanced recognition at the 3D-PSSM (three-dimension position-specific scoring matrix) server (http://www.bmm.icnet.uk/~3dpssm/) (Kelley, MacCallum et al. 2000); (3) the GenTHREADER program at the PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/) (Jones 1999); (4) Sausage (Sequence-structure Alignment Using a Statistical Approach Guided by Experiment) server (http://rsc.anu.edu.au/~drsnag/TheSausageMachine.html) (Huber, Russell et al. 1999); (5) the secondary structure prediction - based fold recognition server, TOPITS (http://www.embl-heidelberg.de/predictprotein/predictprotein.html) (Rost 1995; Rost, Schneider et al. 1997); (6) FFAS (Fold & Function Assignment System) server (http://bioinformatics.ljcrf.edu/FFAS/) (Rychlewski, Jaroszewski et al. 2000); (7) sequence-structure homology recognition server that uses environment-specific substitution tables and structure-dependent gap penalties, FUGUE, at http://www-cryst.bioc.cam.ac.uk/~fugue/ (Shi, Blundell et al. 2001).

2.3 RESULTS

2.3.1 PSI-BLAST Searches

GyrA and ParC homologues were found in PSI-BLAST searches initiated from the C-terminal fragment of *M. genitalium* GyrA as described in Materials and Methods. Inspection of local alignments generated by PSI-BLAST revealed the presence of multiple high scoring pairs (HSPs) for as many as 80% of the found homologues, indicating the presence of sequence repeats. In other words, the same segment of the query sequence was aligned to several different segments in the same subject sequence with reliably high E-values (below 0.02). Multiple alignment analysis established the presence of 6 sequence repeats in GyrA and ParC C-terminal domains (Figure 2.2).

PSI-BLAST iterations initiated from most of the GyrA and ParC sequences converged within the type II topoisomerase family and did not result in structural predictions. However, the 3rd iteration with the query gi|544464, which is annotated as *Fibrobacter succinogenes* GyrA, yielded one non-topoisomerase sequence with an E-value of 0.017 (bit score 40, NCBI nr database, September 2001, 751,829 sequences, 239,148,880 total letters). This sequence, human cell cycle regulatory protein (gi|87057, residues 80-206), is a variant of human RCC1, which has a known three-dimensional structure (gi|4389390, PDB entry 1a12) (Renault, Nassar et al. 1998; Renault, Kuhlmann et al. 2001) and can offer a fold prediction for the C-terminal fragment of GyrA/ParC. RCC1 folds as a 7-bladed β-propeller, with blades being coded by sequence repeats. Each blade is composed of 4 antiparallel β-strands. No sequences from other families were found with significant E-values.

2.3.2 Secondary Structure Predictions And Fold Recognition

JPRED2 secondary structure predictions (Cuff and Barton 2000) obtained for several gyrase sequences strongly suggest that they are all-beta proteins (Figure 2.2). Most of the β-strands were predicted with high confidence level (PHD confidence 7-9, Figure 2.2). SAM-T99 secondary structure prediction yielded similar results (Figure 2.2). β-Strands 5 residues long on average were predicted along the sequence with spacing of about 2-20 residues between them. Secondary structure predictions were similar for the sequence repeats with the consensus prediction of 4 β-strands per repeat (Figure 2.2). The secondary structure prediction for RCC1 sequences were similar and in agreement with the crystal structure of RCC1. Furthermore, the secondary structure predictions show an excellent correspondence between the GyrA/ParC C-terminal domain and RCC1 families.

The consensus fold recognition method of Fischer that combines sequence, structural, and evolutionary information (Fischer 2000) was applied to several topoisomerase II sequences. 7-bladed or 6-bladed β-propellers were consistently found as the top scoring proteins. For instance, the top three fold recognition hits for gyrase gi|121882 are: a theoretical model of human nidogen ywtd β-propeller domain (PDB entry 1NDX, score 17.8); C-terminal WD40 domain of tup1 (PDB entry 1ERJ, score 17.4); and phytase from *Bacillus amyloliquefaciens* (PDB entry 1CVM, score 13.0). There is a substantial gap in the consensus scores between the top three hits and the fourth one with the score of 5.7, which suggests that no other known fold "fits" the gyrase sequence well. In the results from 3D-PSSM, 6-bladed or 7-bladed β-propellers were the top scoring protein folds with 0.05-0.5 PSSM E-values and 90-50% certainty. Furthermore, query gi|1346229 found RCC1 at PSSM E-value of 0.533, with 50% certainty. The results from FFAS also showed 7- or 6-bladed propeller as top hits. gi|68494 found RCC1 as the second hit with E-value of 13.4 and Z-score of 6.02. gi|121882 found RCC1 as the third hits with E-value of 31.2 and Z-score 5.76. FUGUE also found 7- or 4-bladed β-propellers as top hits, but failed to find RCC1. Sausage found β-propellers and antiparallel β-sheet proteins as top hits for the majority of the query sequences. For gi|1346229, it found RCC1 as the top hit with a score of 3.31. TOPITS and

GenTHREADER did not find β-propellers, other mostly β-sheet proteins were the top hits with marginal statistics.

2.3.3 Multiple Sequence Alignment

PSI-BLAST searches demonstrated that sequence repeats in GyrA/ParC are more similar to each other than to repeats in other proteins. Thus GyrA/ParC repeats should be more easy to align with each other. RCC1 family was the only group that displayed statistically supported sequence similarity (PSI-BLAST E-value of 0.017, 12%-29% identity) to GyrA/ParC repeats. Therefore we selected RCC1 for more detailed analysis.

To probe further potential homology between the GyrA/ParC C-terminal domain and RCC1, a multiple sequence alignment was constructed (Figure 2.2). The alignment confirmed the presence of 6 repeats in GyrA/ParC sequences. Each repeat was predicted to contain 4 β-strands (A to D). Loops were relatively short (2-6 residues) between all but two β-strands. Only between β-strands C and D loops were longer (typically about 15 residues). The alignment revealed conservation of hydrophobic residues in β-strands, conserved positively charged residues in β-strand C, and a pair of conserved small residues (typically glycines) in each repeat (Figure 2.2).

The alignment of the RCC1 family was constructed independently and showed 7 sequence repeats with 4 predicted β-strands in each repeat in agreement with the crystal structure of human RCC1. The alignments of GyrA/ParC and RCC1 were merged on the basis of PSI-BLAST local alignments that superimposed the long loops between the strands C and D (Figure 2.2). Such alignment results in a different placement of the Velcro of the propeller in GyrA/ParC and RCC1. In RCC1, Velcro is between the strands B and C. GyrA/ParC are predicted to have a Velcro between A and B. To obtain additional support for the register of β-strands between GyrA/ParC and RCC1, average hydrophobicities were calculated for each β-strand in GyrA/ParC and RCC1 (Table 2.1). Comparison of the 4 hydrophobicity values confirm the alignment of β-strands and thus Velcro placement in GyrA/ParC.

2.4 DISCUSSION

2.4.1 Validity Of The Fold Prediction

The results of PSI-BLAST searches, secondary structure predictions, fold recognition and multiple alignment analysis allow us to deduce the fold of the GyrA/ParC C-terminal fragment. The presence of 6 sequence repeats with 4 predicted β-strands each (Figure 2.2), the PSI-BLAST hit to RCC1, and the detection of propeller folds with threading method strongly argue that the GyrA/ParC domain adopts the 6-bladed β-propeller structure.

Proper alignment of the GyrA/ParC sequences with the RCC1 structure is challenging because of the low level of sequence similarity. Most importantly, corresponding β-strands in GyrA/ParC and RCC1 should be found and correctly aligned. Due to repetitive sequences in GyrA/ParC and the hydrophobic character of β-strands, it is potentially possible to miss the register of β-strands and to align a β-strand in GyrA/ParC to a non-equivalent β-strand in RCC1. For instance, the inner β-strand of the propeller blade may be incorrectly aligned with the outer β-strand.

Three lines of evidence support the alignment presented in Figure 2.2. First, it matches pairwise alignments produced by automatic tools such as PSI-BLAST and the fold recognition method of Fischer. Second, the longest loop between the strands (C and D) in GyrA/ParC is aligned with the longest loop between the strands in RCC1. Third, and most importantly, hydrophobicity analysis of β-strands reveals correspondence in patterns between GyrA/ParC and RCC1 (Table 2.1). Each blade of the propeller is composed of 4 β-strands (A, B, C, D). Since these β-strands are placed at non-equivalent positions in the overall circular structure of the propeller (Figure 2.3a), average hydrophobicities of these 4 β-strands differ. The β-strand D is the outermost strand, and it is the most exposed. Thus the β-strand D is expected to be the most hydrophilic. The β-strand A is the innermost strand located along the central shaft of the propeller. The shaft of the propeller contains water molecules and thus the β-strand A is not expected to be the most hydrophobic. The β-strand B is the one with the highest hydrophobicity (Table 2.1). Excellent fulfillment of these tendencies in GyrA/ParC and RCC1 families strongly supports the alignment on Figure 2.2.

2.4.2 Structural Differences Between Gyra/Parc And RCC1

Typically, homology-based predictions can deduce only similarities between the query and its homologue with experimentally determined structure. The differences are more challenging to predict. Some differences may be wrongly missed and similarities be falsely predicted instead. Such bias is more likely to occur at very low sequence similarity levels when homology is remote. This is the case with GyrA/ParC-RCC1 superfamily. Here we argue that the two most important differences between GyrA/ParC and RCC1 can be predicted.

First, GyrA/ParC should fold as a 6-stranded propeller rather than a 7-stranded propeller as RCC1. This simply follows from the fact that only 6 sequence repeats can be detected in GyrA/ParC sequences. The sequences outside the 6-repeat fragment either belong to the domain of determined structure (N-terminal to the first repeat) or lack clearly predicted β-strands (the extreme C-terminal region). Additionally, the fragment of GyrA that corresponds exactly to the 6 repeats is naturally expressed in *Borrelia burgdorferi* (see discussion below) (Knight and Samuels 1999). Homology between propellers that display different number of blades have been reported before (Wolf, Brenner et al. 1999) and therefore is not surprising.

Second, the Velcro position should differ between GyrA/ParC and RCC1 propellers (Figure 2.2). In RCC1, the first blade starts from β-strand C and the last blade ends with the β-strand B. Thus, one half of the first blade is made from the N-terminal β-strands of the protein, and the other half is made from the C-terminal β-strands (2+2 Velcro). Such an assembly is favorable for the stabilization of the circular arrangement of blades. Since the first repeat of GyrA/ParC starts from the β-strand B and the last repeat ends with the β-strand A, the stabilization of the propeller circular arrangement is probably achieved by a 1+3 rather than 2+2 combination of β strands. This 1+3 Velcro is known for other propellers such as methylamine dehydrogenase (PDB entry 2BBK), nitrite reductase (PDB entry 1NIR) and tachylectin-2 (PDB entry 1TL2), however 2+2 Velcro of RCC1 is apparently unique (Paoli 2001).

2.4.3 Functional Implications

The GyrA/ParC C-terminal domain remains the longest sequence segment of topoisomerase II without available structural information. Therefore the function of this domain is not fully understood despite some effort in this direction. The structure prediction presented here and homology of the GyrA/ParC domain with the RCC1 protein have several functional implications. The RCC1 molecule functions as a protein-binding and a DNA-binding module. One side of the propeller accommodates a protein (Ran) binding site, and the Ran-RCC1 complex structure is available (Renault, Kuhlmann et al. 2001). It is believed that the opposite side of the propeller is involved in interactions with DNA. Available experimental information about GyrA/ParC C-terminal domain suggests similar properties. Being expressed separately, the GyrA domain can associate with the rest of the A subunit, thus possessing a protein binding site. GyrA domain lacks catalytic activity, but binds DNA in a non-sequence specific manner, therefore it should have a nucleic acid binding site.

It has been demonstrated that *Borrelia burgdorferi* expresses a 34 kDa fragment translated from an abundant transcript initiated within the GyrA coding region (Knight and Samuels 1999). This fragment corresponds exactly to the 6 blades of the predicted β-propeller structure, starting from the strand B and ending with the strand A. *Borrelia burgdorferi* gives a unique example, for prokaryotes, of constitutive expression of two proteins, one being a fragment of another, from the same open reading frame. It has been shown that a naturally synthesized transcript abundant in *Borrelia burgdorferi* corresponding to the predicted β-propeller functions as a non-specific DNA-binding protein, forming higher-order nucleoprotein complexes (Knight and Samuels 1999).

Our prediction allows researchers to visualize the distribution of residues in space for the C-terminal domain of GyrA/ParC, despite its unsolved structure. The structural diagram of the C-terminal domain of RCC1 is shown in Figure 2.3a. We predict protein- and DNA-binding surfaces in GyrA/ParC to be similar to the ones in RCC1 (Figure 2.3bc). One way to visualize sequence properties on a structure is to use conservation mapping (Pei and Grishin 2001). The conservation in the blade-to-blade alignments of all available sequences of GyrA/ParC and RCC1 is mapped onto the structure of the 3rd blade in RCC1 (Figure 2.3bc). Similarities in conservation between GyrA/ParC and RCC1 include mainly small residues (C,A,P,S,T) in loops. These residues bear potential structural importance. The most pronounced difference in conservation patterns of GyrA/ParC and RCC1 is due to the presence of a conserved residue stretch closer to the N-terminus of the β-strand C in GyrA/ParC. These conserved residues are mainly positively charged (shown in blue in Figure 2.2) and could potentially contribute to a DNA-binding site in GyrA/ParC.

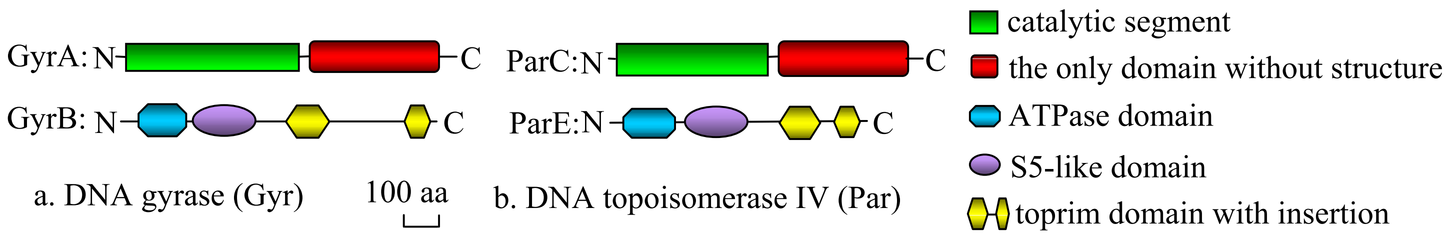
2.4.4 Prediction Confirmation

Our prediction was made in March 2001 and was published in May 2002 (Qi, Pei et al. 2002). Two years later, experimentally determined structures for GyrA and ParC C-terminal domains were published (Corbett, Shultzaberger et al. 2004; Hsieh, Farh et al. 2004). Figure 2.4 shows the structure diagram of GyrA C-terminal domain determined by experiment (Corbett, Shultzaberger et al. 2004), which is a 6-bladed β-propeller with 4 β–strands in each blade as predicted and also with a 1+3 Velcro. As summarized in Table 2.2, the experimentally determined GyrA/ParC C-terminal domain structures confirmed our structural fold prediction; while exhibiting a novel blade topology different from the canonical one.

2.5 CONCLUSIONS

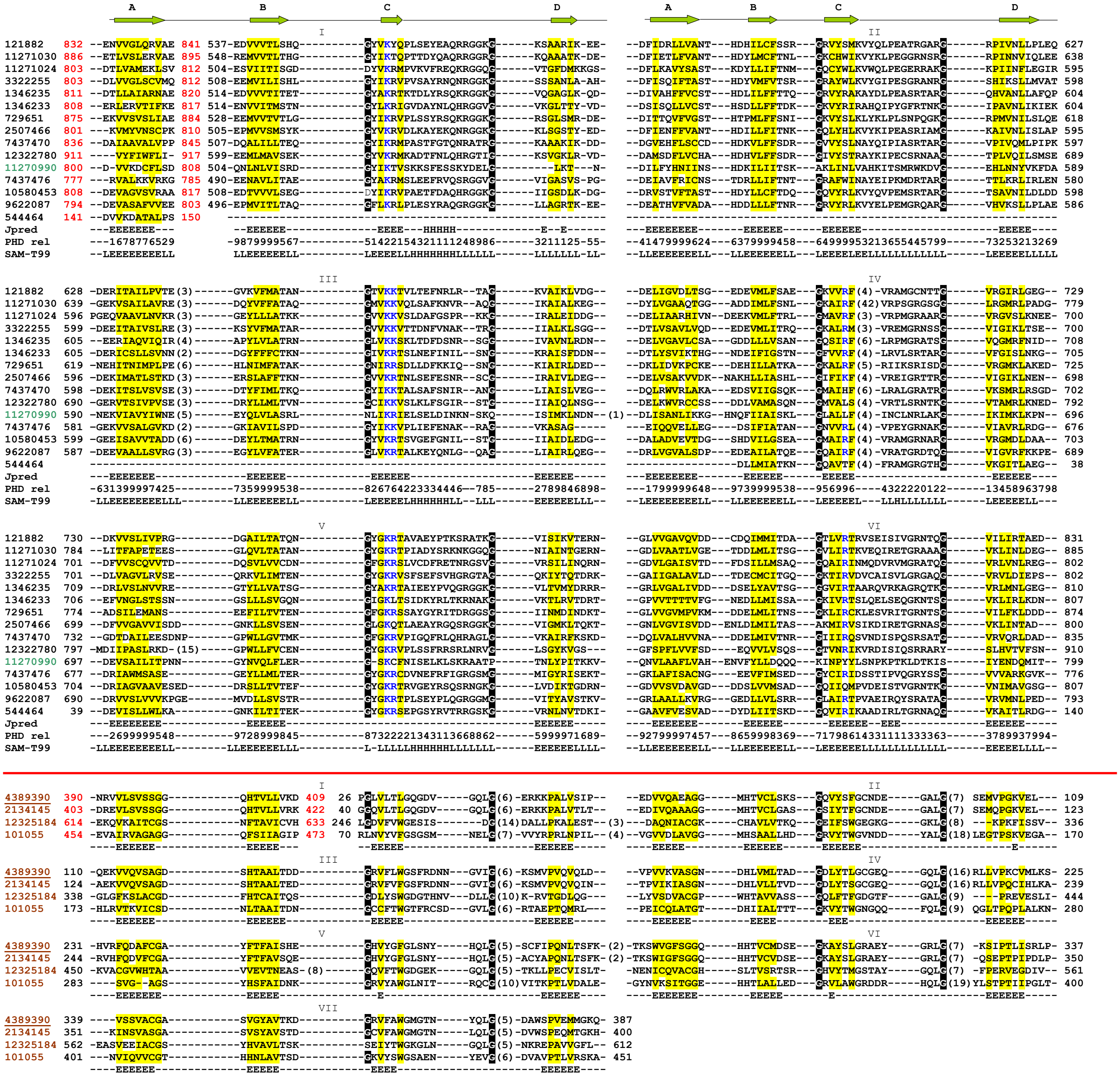
In this case study of structure modeling and prediction, we have detected sequence similarity between C-terminal domain of GyrA/ParC and regulator of chromosome condensation (RCC1) and infered homology relationship between them. The results of hydrophobicity analysis, secondary structure prediction and fold recognition all support the inferred homology relationship. Based on these extensive sequence and structure analysis, the C-terminal domain of GyrA/ParC has been predicted to have a 6-bladed b-propeller structure with 4 b–strands in each blade. Experimentally determined GyrA/ParC C-terminal domain structures confirm the structural fold prediction, while exhibit a novel blade topology different from the canonical one.

Figure 2.1 Domain compositions of gyrase and topoisomerase IV



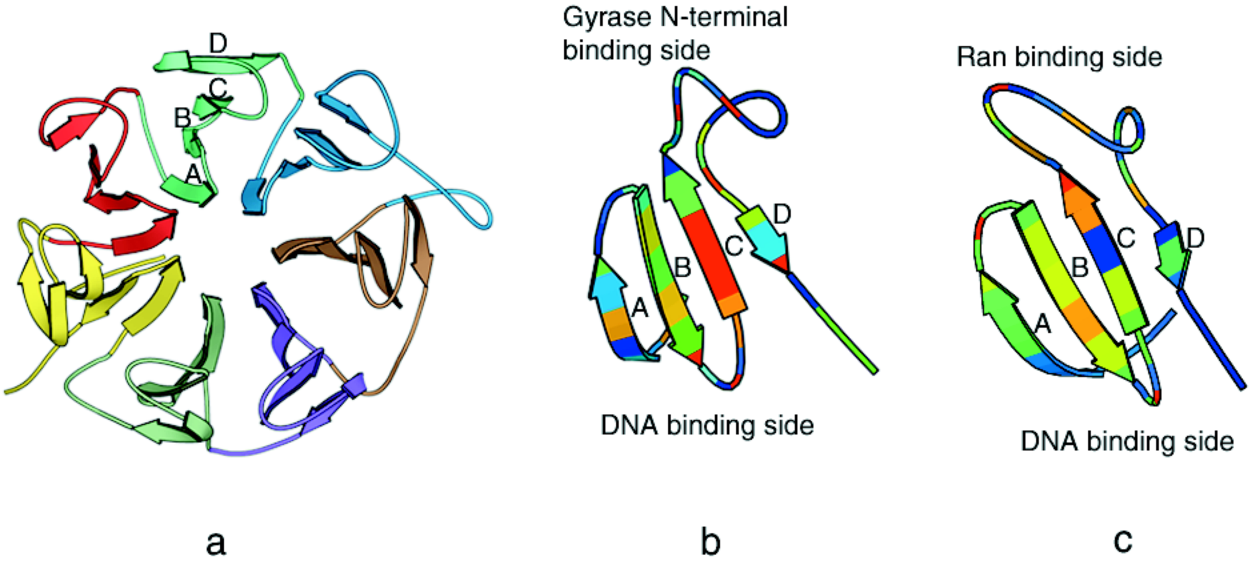
Domain compositions of (a) gyrase (topoisomerase II Gyr) and (b) topoisomerase IV (Par). Sequences shown are all from *Escherichia coli.*

Figure 2.2 Multiple sequence alignment of C-terminal of GyrA/RCC1 domain



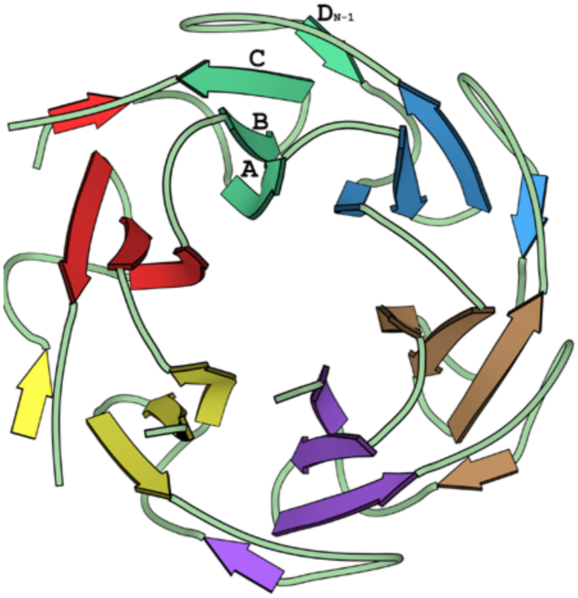
Multiple sequence alignment of C-terminal of gyrase subunit A/RCC1 domain. Each sequence is labeled by its NCBI gene identification (gi) number. The gi numbers of GyrA/ParC C-terminal domains and the gi numbers of RCC1 domains are in black and brown, respectively. The gi number of the topoisomerase IV subunit A sequence (gi|11270990) is in green. The gi number of the sequence with known structure is underlined (gi|4389390; PDB entry 1A12, chain A). The alignment is arranged in such a way that each row of sequences contains two blades. The blades are numbered from above using Roman numbers. The last β-strand A of GyrA/ParC C-terminal domain was placed in front of the first β-strand B of blade I to complete the blade. The sequences of RCC1 were rearranged in the same manner. The first and last residue numbers of each row of sequences are indicated. The first and last residue numbers of the rearranged C-terminal segments are marked in red. Long insertions in loop regions are not shown but with the omitted residues numbers in parentheses. Uncharged residues at mainly hydrophobic positions are shaded yellow. The conserved glycine residues are shown in white on black background. Conserved positively charged residues in β-strand C are shown in blue. The JPRED secondary structure prediction results are the first lines shown below each row of the alignment. The PHD prediction confidence values of every position for GyrA/ParC are shown on the second line under the predictions. The third lines under the GyrA/ParC alignment are the secondary prediction results from SAM-T99. The diagram of the secondary structure elements in each blade, according to the RCC1 X-ray structure, is shown at the top of the figure. Species names: gi|121882, *Escherichia coli*; gi|11271030, *Neisseria meningitides*; gi|11271024, *Chlamydia muridarum*; gi|3322255, *Treponema pallidum*; gi|1346235, *Mycobacterium tuberculosis*; gi|1346233, *Mycoplasma genitalium*; gi|729651, *Rickettsia prowazekii*; gi|2507466, *Helicobacter pylori*; gi|7437470, *Synechocystis sp*; gi|12322780, *Arabidopsis thaliana*; gi|11270990, *Ureaplasma urealyticum*; gi|7437476, *Archaeoglobus fulgidus*; gi|10580453, *Halobacterium sp.*; gi|9622087, *Thermus thermophilus*; gi|544464, *Fibrobacter succinogenes*; gi|4389390, *Homo sapiens*; gi|2134145, *African clawed frog*; gi|12325184, *Arabidopsis thaliana*; gi|101055, *Schizosaccharomyces pombe*.

Figure 2.3 Structure Diagrams and Conservation Mapping of RCC1 and GyrA/ParC-CTD



(a) The structural diagram of RCC1, PDB entry 1A12 chain A. Each blade is shown in a different color and β-strands in the third blade are labeled. Sequence conservation in (b) GyrA/ParC and (c) RCC1 mapped onto the structure of the third blade in RCC1 are rainbow colored from low conservation (dark blue) to high conservation (red).

Figure 2.4 Experimentally Determined Structure Diagram of GyrA-CTD

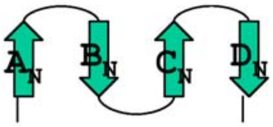
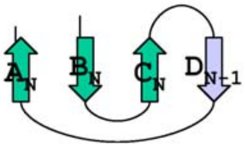


Structure diagram of experimentally determined spatial structure of GyrA C-terminal domain (PDB accession number: 1SUU)

Table 2.1 Average hydrophobicity of beta-strands in GyrA/ParC and RCC1

	GyrA/ParC	RCC1
Strand A	0.21	0.256
Strand B	0.37	0.34
Strand C	0.17	0.29
Strand D	0.085	-0.13

Table 2.2 Comparison Between Prediction and Experimental Data

	Prediction	Experimental Data*
Structural Fold	6-bladed β-propeller with 4 β–strands in each blade	6-bladed β-propeller with 4 β–strands in each blade
1+3 Velcro	Yes	Yes
Relative positions of β-	ABCD (Inner-most to outer-most)	ABCD (Inner-most to outer-most)
Blade Topology	 canonical	 novel

* Based on PDB 1SUU and 1WP5.

CHAPTER 3:
Structural Classification of Thioredoxin-like Fold Proteins

3.1 INTRODUCTION

3.1.1 Background

A systematic comparison of the three major structure classifications (SCOP, CATH, DaliDD) shows many discrepancies, even at the fold group level (Hadley and Jones 1999). These discrepancies create obstacles for homology inference and modeling, evolutionary studies and genome annotation. One major source of the inconsistencies stems from the concept of fold definition. Structural fold concept is a perception of a researcher and thus is intrinsically subjective. The definition of a protein fold is therefore somewhat arbitrary. For example, it is difficult to define and to distinguish folds of regular-layered architectures, especially α/β sandwiches. Their β-sheets take up a large proportion of the structure and are similar due to hydrogen-bonding constraints, and the differences between structures could be only a few secondary structure elements (Orengo, Flores et al. 1993; Orengo, Michie et al. 1997). In an effort to understand and to clarify fold definitions for proteins with α/β sandwich architectures, we start from a large and diverse protein group, namely thioredoxin-like proteins.

Thioredoxin is an important redox protein that is present in every organism. Together with thioredoxin reductase and peroxiredoxin, thioredoxin regulates the cellular reduction/oxidation status as well as various important cellular functions, such as oxidative stress defense, cell proliferation, signal transduction, and transcription regulation (Nakamura, Nakamura et al. 1997; Arner and Holmgren 2000; Yamawaki, Haendeler et al. 2003; Das 2004; Kontou, Will et al. 2004). Extensive studies have been done on Thioredoxin (Holmgren 1995; Nakamura, Nakamura et al. 1997; Arner and Holmgren 2000; Yamawaki, Haendeler et al. 2003; Kontou, Will et al. 2004). Consequently, a large number of X-ray and NMR structures are available for thioredoxin and related proteins, rendering their classification necessary.

3.1.2 Objective

Figure 3.1c shows the structure of a human thioredoxin, which is a 3-layer α/β/α sandwich with the central β-sheet formed by 5 β-strands flanked by two α-helices on each side. Many proteins important for cellular thiol-redox pathways, such as glutaredoxin, protein disulfide isomerase (PDI) and oxidase (DsbA), and glutathione S-transferase (GST), are homologous to thioredoxin and have similar structures. However, many of these classical thioredoxin-like proteins do not contain α-helix α0' and β-strand β0', and some do not contain α-helix α3' (Figure 3.1c). To generate a consistent and inclusive definition of the thioredoxin-like fold, we use the structure consensus of thioredoxins and the classical thioredoxin-like proteins that are undoubtedly homologs to each other, and only include those secondary structure elements and interactions that are present in all these homologs (Figure 3.1a). Interestingly, a circularly permuted DsbA protein exists as a result of a protein engineering experiment that is structurally stable and functionally active (Hennecke, Sebbel et al. 1999). As homologous proteins can evolve to have different circular permutations (e.g., DNA methyltransferases (Jeltsch 1999)), we decide not to limit our fold group definition to identical topology, but to consider all potential circular permutations of the thioredoxin-like fold.

We employ this definition of the thioredoxin-like fold to query the PDB database using a protein structure motif search program (unpublished). Identified thioredoxin-like protein domains are divided into eleven evolutionary families based on combined sequence, structural and functional evidence for homology. Analysis of the protein-ligand structure complexes reveals two major active site locations for thioredoxin-like proteins. During the course of analysis, we also encountered proteins with structural similarity to thioredoxin that should not belong to the thioredoxin-like fold group. Such examples are shown and discussed to illustrate our approach to fold definition.

3.2 MATERIALS AND METHODS

3.2.1 Structural Motif Search For Thioredoxin-Like Protein Domains

We used a structure motif search program (unpublished) that was under development in our lab. Briefly, the program generated a database of PDB structures (19,558 structures, July 2003), in which each structure was represented by a Secondary Structure Element Interaction Matrix describing the interactions (parallel or anti-parallel), hydrogen-bonding and chirality between the secondary structure elements of the PDB structure. The structure consensus of the classical thioredoxin-like proteins (thioredoxin, glutaredoxin, protein disulfide isomerases (PDI), disulfide bond oxidase (DsbA), glutathione S-transferase (GST), glutathione peroxidase and their close homologs) (Martin 1995) was represented as a query matrix. The query matrix (Figure 3.1b) specified the number and types of secondary structure elements in the thioredoxin motif, the hydrogen-bonding and parallel or anti-parallel relationships between the four β-strands, and the chirality between consecutive secondary structures. We then used our structure motif search program to search the database of Secondary Structure Element Interaction Matrices of every PDB structure and to output the structures containing submatrix matching the query matrix. Six query matrices characterizing six possible circular permutations of the thioredoxin motif were constructed and searched for. False positives were removed by visual inspection. Proteins were considered to contain the thioredoxin-motif only when the thioredoxin motif formed the structural core of the protein domain (see “Structural analogs” section for details).

3.2.2 Sequence-Based Classification Of The Thioredoxin-Like Protein Domains

The thioredoxin motif-containing protein domains retrieved as described above were subsequently grouped into evolutionary families using a combined sequence, structural and functional analysis.

We used four methods to search for sequence similarities between the thioredoxin motif-containing domains and all PDB proteins: gapped BLAST (Altschul, Gish et al. 1990; Altschul, Madden et al. 1997), PSI-BLAST (Altschul, Madden et al. 1997; Schaffer, Aravind et al. 2001), RPS-BLAST (Marchler-Bauer, Panchenko et al. 2002) and COMPASS (Sadreyev and Grishin 2003), each of which uses a query sequence or profile to search a database of sequences or profiles. A query sequence was the sequence of every thioredoxin-motif containing domain. A query profile was generated by running a query sequence against the nr database (1,479,768 sequences, 476,959,297 total letters, Aug 2003) using PSI-BLAST for up to 5 iterations with an inclusion E-value cutoff of 0.005. The database of PDB sequences contained sequences of PDB chains (49,319 sequences, 10,645,968 total letters, Aug 2003). The database of domain profiles contained the profiles of representative protein domains in the PDB. We used the SCOP v1.63 domain definitions for this purpose. The representative SCOP v1.63 domain sequences with less than 40% sequence identity to each other (5,224 domains) were downloaded from Astral (Brenner, Koehl et al. 2000; Chandonia, Walker et al. 2002). A profile for each representative domain sequence was then generated in the same way as we generated a query profile. We searched each query sequence in the database of PDB sequences using Gapped BLAST (Altschul, Gish et al. 1990), each query profile in the database of PDB sequences using PSI-BLAST (Altschul, Madden et al. 1997; Schaffer, Aravind et al. 2001), each query sequence in the database of domain profiles using RPS-BLAST (Marchler-Bauer, Panchenko et al. 2002), and each query profile in the database of domain profiles using COMPASS (Sadreyev and Grishin 2003). Sequence analyses were based on the search results of the four methods. We also inspected each hit with an E-value up to 10 so that we would not miss a potential homolog that has a signature sequence motif but with a less significant E-value.

3.2.3 Structure And Function Based Classifications

For structure analysis, 723 thioredoxin motif-containing protein domains were first clustered according to their sequence identities using the program BLASTCLUST (I. Dondoshansky and Y. Wolf, unpublished; ftp://ftp.ncbi.nih.gov/blast/) at a sequence identity threshold of 50% and length coverage of 90%. A representative structure for each cluster was selected based on the quality of the structure (resolution, R factor value, solved date for NMR structures) and the presence of ligands or substrate analogs. All structure analyses were done on this set of the representative domain structures. The representative structures were aligned in an all-against-all manner using the program DaliLite and were further clustered by a Dali Z-score cutoff of 5. The representative structures were visualized in the INSIGHT II package (MSI) and superimposed by aligning structurally equivalent residues. A structure-based multiple sequence alignment of all 90 representative structures was constructed manually taking into account alignments made by DaliLite (Holm and Park 2000), Mammoth (Ortiz, Strauss et al. 2002), CE (Shindyalov and Bourne 1998), PSI-BLAST (Altschul, Madden et al. 1997; Schaffer, Aravind et al. 2001) and RPS-BLAST (Marchler-Bauer, Panchenko et al. 2002). The structural alignment was further filtered by sequence identities in the aligned regions and the final alignment contained proteins that had less than 50% sequence identity to each other. The ligands or substrate analogs and active site residues were also visualized in INSIGHT II and locations of active sites were compared.

3.3 OVERALL FOLD DESCRIPTION

3.3.1 Thioredoxin-like fold

Many proteins important for cellular thiol-redox pathways, such as thioredoxin, glutaredoxin, glutathione S-transferase (GST), protein disulfide bond isomerase (PDI), are known to adopt the thioredoxin-like fold (Martin 1995). In both SCOP and CATH, the thioredoxin/glutaredoxin fold is described as a 3-layer $\alpha/\beta/\alpha$ sandwich. As shown in Figure

3.1c, thioredoxin is a 3-layer sandwich with a central β-sheet flanked by two α-helices on each side. However, the N-terminal α-helix α0' is absent in many classical thioredoxin-like fold proteins, such as GST, bacterial glutaredoxin, and archaeon PDI. α-Helix α3' is also not conserved in the thioredoxin homologs. For instance, the N-terminal domain of a bacterial alkyl hydroperoxide reductase subunit F (1hyuA1), which is a close homolog of PDI, has only a short loop connecting β2 and β3 in the place of the α-helix α3' (Figure 3.2). In addition, phosducin (1a0rP), a homolog of thioredoxin, has only a loop with turns in the place of the α-helix α3' (Figure 3.2). In many proteins that do have α-helices at the α3' position, these α-helices are irregular, kinked or appear as separated short helical turns. Based on these observations, the first α-layer of the thioredoxin fold is not conserved in all thioredoxin homologs. Since the fold definition should include only the core secondary structural elements that are present in the majority of homologs, we define the thioredoxin-like fold as a 2-layer α/β sandwich with the βαβββα secondary structure pattern. The four β-strands ordering 2134 form a mixed β-sheet with the third β-strand anti-parallel to the rest, and the two α-helices pack against the β-sheet on one side (Figure 3.1a). The N-terminal half of the fold is a right-handed βαβ unit. This unit is connected through a loop to the C-terminal half of the fold, which is a β-hairpin followed by an α-helix and the chirality of this ββα unit is left-handed. Consequently, the chiralities between secondary structure elements β4, α2, β1, and α2, β1, α1 are both right-handed.

Applying this definition, we searched for all potential thioredoxin-like protein domains in the entire PDB database using the structure motif search program under development in our lab. Found proteins containing the βαβββα unit with the thioredoxin-like interactions (see materials and methods and Figure 3.1) were visually inspected to ensure that the six elements form the structural core (see “Structural analogs” section for clarification) of the protein domains. Altogether 723 protein domains were identified as thioredoxin-like fold proteins. They were unified into the thioredoxin-like fold group and divided into evolutionary families. A structure-based multiple sequence alignment of 90 representative thioredoxin-like fold protein domains was manually constructed (Figure 3.2). From this alignment, we see that some thioredoxin-like proteins have insertions of secondary structure elements into the common structural motif. A number of proteins from four families possess the α-helix α3'. Proteins from other four families have an extra αβ unit inserted between the β-strands β2 and β3, extending the central β-sheet to be formed by 5 β-strands.

3.3.2 Circular Permutations

The protein domains that we unified into the thioredoxin-like fold group represent different circular permutations of the thioredoxin-like motif. A circular permutation of a structural motif can be visualized as an imaginary "ligation" of the N- and C- termini followed by an imaginary "cleavage" at a loop region of the motif to create different termini. Except when specifically mentioned, we use the phrase "circular permutation" only to indicate this kind of geometric relationship between structures and not to imply evolutionary events. It has been documented, however, that circular permutations occur in nature as evolutionary scenarios and represent a mechanism of potential fold change in evolution (Ponting and Russell 1995; Gong, O'Gara et al. 1997; Jeltsch 1999; Bujnicki 2002). Since proteins with different circular permutations of a structural motif have essentially the same spatial arrangement of secondary structure elements, the same side-chain packing interactions and may be homologous, grouping them together into the same fold group for further comparative analysis could help us to better understand protein folding and sequence-structure-function relationships and potential evolutionary connections. We can use the structure-based multiple sequence alignment to study the sequence similarities between proteins with different circular permutations. Such potential similarities are obscured if the proteins are classified in different fold groups or even different structural classes.

Since the thioredoxin-like motif contains six secondary structure elements, six types of circular permutations are theoretically possible by placing the termini before each secondary structure element. However, only four types of circular permutations were seen in the PDB database (Figure 3.1a). No proteins are present with the termini positioned between β1-α1 or α1-β2, suggesting that β1α1β2 may be an essential folding or packing unit for the thioredoxin-like fold. This observation agrees with the finding by Salem *et. al.* that βαβ-unit is one of the three most prominent (highly-populated) supersecondary structures (Salem, Hutchinson et al. 1999). However, it is possible that with more structures accumulating in the PDB, circular permutation variants that disrupt the βαβ-unit will appear. Out of the four types of circular permutations we see, type II (β4α2β1α1β2β3; secondary structures are numbered the same as those in the classical thioredoxin-like proteins) is adopted in five families and the other three types are all adopted in two families, respectively (Table 3.1). If we count the number of representative structures, type I (β1α1β2β3β4α2) is the most populated, and type II is the second-most populated type of circular permutation.

3.4 DESCRIPTION OF THIOREDOXIN-LIKE FOLD FAMILIES

We identified 723 protein domains as belonging to the thioredoxin-like fold. We subsequently classified these protein domains into eleven evolutionary families based on inferred homology relationships between them. While we gathered strong support for homology of protein domains within each evolutionary family, we are not drawing any conclusion about the evolutionary relationship between protein domains in different families. Protein domains from different families could simply be analogous to each other. Alternatively, they could share homologous relationship that we were not able to support convincingly, or be mosaics of homologous and analogous pieces. It has been hypothesized that modern protein domains have evolved from combinations of ancient domain segments composed of supersecondary structures, and thioredoxin fold proteins is a possible example of such domain evolution (Lupas, Ponting et al. 2001). Although a detailed analysis of this problem is very challenging and lies beyond the scope of our current study, this evolutionary scenario is plausible. However, we believe that for the proteins within each of our evolutionary families homologous segment spans through the entire common core of the domain. Here we describe the eleven families and discuss their sequence, structural and functional features with evolutionary implications. The representatives of each family are listed in Figure 3.2.

3.4.1 Thioredoxin family

This family includes all the classical thioredoxin-like proteins as well as calsequestrin, phosducin and arsenate reductase, among others. The dithiol-disulfide oxidoreductases, such as thioltransferases and PDI, have a conserved active-site sequence motif Cys-X-X-Cys that is located at the N-terminus of α-helix $\alpha 1$. In addition, a cis-proline residue located at the loop region before $\beta 3$ is conserved and is in spatial proximity to the Cys-X-X-Cys motif (Figure 3.1c). Proteins that form inter-domain disulfide bonds, such as glutathione peroxidases, and proteins that do not form disulfide bonds, such as the N-terminal domain of elongation factor 1-gamma (eEF1gamma), have lost one or both of the conserved Cys residues (Figure 3.2). Nevertheless, they have the same active site locations as the dithiol-disulfide oxidoreductases, and their homology relationships with the dithiol-disulfide oxidoreductases can be inferred from PSI-BLAST and RPS-BLAST results and close structural similarities.

Protein domains in this family have a type I circular permutation except for one disulfide bond oxidase (DsbA, 1un2A, previous PDB ID: 1dyv) that is a type III circular permutation as the result of a protein engineering experiment (Hennecke, Sebbel et al. 1999). Aside from the common structural motif, most thioredoxins and PDIs have the extra α-helices $\alpha 0'$ and $\alpha 3'$ (Figure 3.1c). Glutathione peroxidases and peroxiredoxins have an extra α/β unit inserted between $\beta 2$ and $\beta 3$ and the extra β-strand is hydrogen-bonded with $\beta 2$ (Figure 3.2); DsbAs have an extra β-strand inserted before $\beta 1$ and hydrogen-bonded with $\beta 4$; so they all have a mixed β-sheet of five β-strands.

3.4.2 RTPC small domain family

Similarly to thioredoxins, the small domains of the RNA 3'-terminal phosphate cyclases (RTPC) have the type I circular permutation. However, the β-sheet in this family is much flatter and the β-strands are up to 4 residues longer than those of the thioredoxins. The functional role of the RTPC small domain remains unknown (Palm, Billy et al. 2000).

3.4.3 Ribosomal protein L30e family

Ribosomal protein L30e, eukaryotic peptide chain release factor subunit 1 C-terminal domain (ERF1), and RNA 2'-O ribose methyltransferase N-terminal domain are grouped in this family. Inferred from sequence similarity analyses, ribosomal proteins L30e, L7ae and 15.5 kd RNA binding protein are close homologs (gapped BLAST E-value: 2e-11), while ERF1 and L7ae are more distant (gapped BLAST E-value: 0.009). Gapped BLAST, PSI-BLAST, a RPS-BLAST did not find any hit between the RNA methyltransferase N-terminal domain and L30e with E-value less than 10. However, COMPASS aligned the RNA methyltransferase N-terminal domain (1ipaA) and L30e (1cn8A) at a significant E-value of 5e-05. The COMPASS alignment covers the entire length of both domains and is consistent with the structure-based alignment (Figure 3.2), and we thus consider the RNA methyltransferase N-terminal domain to be a remote homolog of ribosomal protein L30e.

Protein domains in this family have a type II circular permutation, and aside from the permutation, are structurally very similar to the thioredoxin family domains. Archaeon ribosomal protein L30 (1h7mA1) superimposes on the thioredoxin family protein eEF1gamma (1nhyA) with a RMSD of 1.4 Å based on 86 C_α atoms. Furthermore, like thioredoxins and PDIs, proteins in this family also have an extra α-helix at the N-terminus (Figure 3.1d) and a α-helix α3' between β2 and β3 to form a second layer of α-helices, and thus also form a 3-layer α/β/α sandwich. However, we think that presently there is not enough evidence to convincingly support this potential homology between the L30e ribosomal proteins and thioredoxins.

Protein domains in the L30e family interact with their ligands and substrates at the N-terminal ends of the α-helices and nearby regions. The yeast ribosomal protein L30 interacts with the RNA internal loop through the residues located at the N-terminal ends of α-helices $\alpha 1$ and $\alpha 2$ and in the loop region before β-strand $\beta 3$ (Mao, White et al. 1999) (Figure 3.1d).

3.4.4 Tubulin C-terminal domain family

This family includes the C-terminal domains of tubulin α- and β-subunit, cell division protein FtsZ, and dihydroxyacetone kinase subunit K (DhaK). The overall structures of tubulin, FtsZ, and DhaK are similar; all are formed of two domains that have the same relative positions. In all proteins of this family, the N-terminal domains are Rossmann-like nucleotide-binding domains: GTPase for tubulin and FtsZ, and ATPase for DhaK. The C-terminal domains are the thioredoxin-like domains with a type II circular permutation. The C-terminal domain of DhaK has a β-hairpin inserted between $\beta 4$ and $\alpha 2$. The substrate Dha is covalently bound (Siebold, Garcia-Alles et al. 2003) to this β-hairpin. In tubulin, the loop between $\beta 4$ and $\alpha 2$ (Figure 3.1a) also forms a functional site where the ligands, zinc ion and anticancer drug taxol, bind (Lowe, Li et al. 2001).

3.4.5 Bacillus chorismate mutase (BCM) Family

Bacillus and *Thermus* chorismate mutase, hypothetical protein YjgF, and purine regulatory protein YabJ are placed in this family. Simple BLAST results show that *Bacillus* with *Thermus* chorismate mutases and YjgF with YabJ form two clusters of close homologs. Despite the low sequence identity (average 8.6%) between the two groups, their tertiary and quaternary structures are very similar to each other. These proteins are homotrimers; each monomer is a thioredoxin-like domain of type II circular permutation. The three β-sheets from three monomers form a barrel-shaped interface. When looking parallel to the three-fold axis that goes in the direction from the C-terminus to the N-terminus of α-helices $\alpha 1$ and $\alpha 2$, the three β-sheets of proteins in both groups run approximately parallel to the axis with a left-handed twist (Figure 3.3a). The monomers of *Thermus* chorismate mutase (1odeA) and YjgF (1qu9A) are superimposed with a RMSD of 1.7 Å based on 84 C_α atoms, and quaternary structures are superimposed very well. The active site locations are also the same for the two group of proteins, which are at the three clefts between adjacent monomers (Chook, Ke et al. 1993; Chook, Gray et al. 1994) (Figure 3.3a), indicating homology.

3.4.6 MECP synthase family

The quaternary structures of 2C-methyl-D-erythritol-2,4-cyclodiphosphate (MECP) synthases are similar to proteins in the *Bacillus* chorismate mutase (BCM) family. MECP synthases are also homotrimers with each monomer a thioredoxin-like domain of type II circular permutation. However, there are several structural and functional site differences between MECP synthases and BCM family proteins. The monomers of MECP synthases have an extra α-helix between α2 and β1 that is absent in the BCM family proteins. The β-sheets of MECP synthases also run approximately parallel to the three-fold axis but with a right-handed twist instead of a left-handed one, so the monomer cannot superimpose well when the trimers are superimposed with the BCM family members. The active sites of MECP synthases are also located at the clefts between adjacent monomers (Kemp, Bond et al. 2002; Kishida, Wada et al. 2003). However, in MECP synthases, the active site residues are contributed from β2 and α1 of one monomer and β4 and α2 of the adjacent monomer; while in BCM family proteins, the active site residues are contributed from β2 and α1 of one monomer but β3 and β4 of the adjacent monomer. These differences between MECP synthases and BCM family proteins indicate that they may not share a common ancestor. Therefore, we place MECP synthases in a separate family.

3.4.7 PurM N-terminal domain family

The N-terminal domain of the aminoimidazole ribonucleotide synthetase (PurM) is a thioredoxin-like domain of type II circular permutation, with an extra 9-residue α-helix inserted between β4 and α2. PurMs are homodimers, and the two β-sheets from the two thioredoxin-like N-terminal domains form a barrel-shaped dimer interface. The active site of PurM is proposed to be formed by the edge β-strands of the two β-sheets and the C-terminal domain (Li, Kappock et al. 1999).

3.4.8 Cytidine deaminase family

This family includes single domain cytidine deaminase (CDA), two-domain CDA, and cytosine deaminase. The N-terminal domain of the two-domain CDA has a higher sequence identity (29%) to the single domain CDA than to its C-terminal domain (15%), suggesting that the two-domain CDA emerged by an ancient gene duplication event of the one-domain CDA and the C-terminal domain diverged further. In fact, the N-terminal domain of the two-domain CDA, the single domain CDA, and the cytosine deaminase all have two conserved cysteines at the N-terminus of α-helix α1 and a conserved cysteine or histidine at the N-terminus of α-helix α2 that coordinate a catalytic zinc ion (Xiang, Short et al. 1996; Johansson, Mejlhede et al. 2002) (Figure 3.1e & Figure 3.2), while the C-terminal domain of two-domain CDA has lost the zinc coordination and thus the catalytic activity.

All domains in this family have a type III circular permutation (β3β4α2β1α1β2, Figure 3.1e), the same as the engineered DsbA. The loop regions before β3 and between β4-α2 are about 8 residues longer than most domains of the thioredoxin family (Figure 3.2), and they form a cover of the hydrophobic active site. Like glutathione peroxidases, cytosine deaminase has an extra α/β unit inserted after β2 and the extra β-strand is hydrogen-bonded with β2 (Figure 3.2). One-domain CDA and the N-terminal domain of two-domain CDA have an extra β-strand inserted after β2 and is also hydrogen bonded with β2, but it is oppositely oriented compared to the extra β-strand in cytosine deaminase.

3.4.9 AICAR Tfase domain of bifunctional purine biosynthesis enzyme ATIC family

The bifunctional purine biosynthesis enzyme ATIC has two functional parts: the inosine monophosphate cyclohydrolase (IMPCH) part and the 5-aminoimidazole-4-carboxamide-ribonucleotide transfermylase (AICAR Tfase) part. ATIC is a homodimer with each monomer participating in both functional parts (Greasley, Horton et al. 2001). Each AICAR Tfase part of the monomer includes two thioredoxin-like domains that are structurally very similar to each other (RMSD of 1.17 Å based on 118 atoms). The two thioredoxin-like domains in the same polypeptide chain are the result of an ancient gene-duplication event, and thus they are homologous to each other. The two thioredoxin-like domains are of type III circular permutation, and like glutathione peroxidases of the thioredoxin family, each of them have an extra $\alpha\beta$ unit inserted after $\beta2$ (Figure 3.3b & Figure 3.2). The second thioredoxin-like domain has an insertion of a small helical domain between $\alpha2$ and $\beta1$. AICAR Tfase has two active sites; each is located between the first thioredoxin-like domain of one monomer and the second thioredoxin-like domain of the other monomer. Our analysis shows that the two homologous thioredoxin-like domains possess different active site locations (Figure 3.3b & section 3.5.1).

3.4.10 Phospholipase D family

This family includes phospholipase D, bacterial nuclease Nuc, and tyrosyl-DNA phosphodiesterase (TDP1). Phospholipase D and Nuc are inferred as close homologs based on RPS- and PSI-BLAST results (RPS-BLAST E-value: 5e-14), while TDP1 was previously shown by Interthal *etc.* (Interthal, Pouliot et al. 2001) to be homologous to phospholipase D and Nuc based on the presence of the conserved HK motif (Figure 3.2) and similar reaction mechanism. Both phospholipase D and TDP1 contain two duplicated thioredoxin-like domains of type IV circular permutation ($\alpha2\beta1\alpha1\beta2\beta3\beta4$). Nuc only contains one such domain, but it is a homodimer and the two monomers are arranged in the same way as the two domains in phospholipase D and TDP1. Like glutathione peroxidases of the thioredoxin family, all protein domains in this family have an extra αβ unit inserted after β2 (Figure 3.2).

3.4.11 gp5 domain A family

Domain A of the major capsid protein gp5 is a thioredoxin-like domain of type IV circular permutation. Protein gp5 is the assembly subunit of the double-strand DNA bacteriophage HK97 capsid (Wikoff, Liljas et al. 2000). Each capsid asymmetric unit is a hexamer or a pentamer of gp5. Other domains of gp5s, domain P, E-loop, and N-arm form a hexagon or a pentagon, and domains A of gp5s form a cover of the space inside the polygon. A 22-residue long insertion between α-helix α1 and β-strand β2 pushes the C-terminus of α1 up and makes α1 almost perpendicular to the β-sheet instead of being parallel to it (Figure 3.1f). This arrangement renders α1 anti-parallel to α-helix α2 of the neighboring gp5 domain A. α1 and α2 of adjacent domains A form electrostatic and hydrophobic interactions in between to stabilize the cover of the polygon.

3.5 DISCUSSION

3.5.1 Analysis of active site locations

Proteins containing the thioredoxin-like domains are involved in a wide variety of biological functions and pathways, including intracellular transport and cell division, signal transduction, pyrimidine salvage pathway, phospholipid metabolism, and biosynthesis of purine, aromatic amino acid and proteins. The thioredoxin-like protein domains can bind and/or catalyze different ligands and substrates such as nucleic acids (RNA and DNA), proteins, peptides, and small metabolites. 3D structure complexes of the protein domains with their ligands or substrate analogs are available for all thioredoxin-like families except the RTPC small domain family. We analyzed the ligands or substrates binding sites of the ten thioredoxin-like fold group families and found two major types of active site locations for the thioredoxin-like protein domains.

In many proteins, active site (type i location) is placed at the N-terminal ends of the α-helices or nearby loop regions, i.e., the binding or catalytic residues are located on the loops connecting β1-α1, β2-β3, β4-α2, or at the N-termini of the α-helices α1/α2 (Figure 3.1a). This type of active site location is adopted by protein domains in five different families that encompass all four circular permutations (Table 3.1). Since protein domains with this active site location belong to different evolutionary families, the similarity in the active site placement may be the result of convergent evolution and is probably caused by physico-chemical constraints such as the helix dipoles of α1 and α2.

Another common placement of the active site (type ii location) is along the edges of the β- sheet, i.e. the binding or catalytic residues are located on the edge β-strands (β2, β4) of the β-sheet or on the sides of α-helices α1 and α2 that are facing opposite from each other. This type of active site location is adopted by protein domains in four different families (Table 3.1). Proteins in three of the families (*Bacillus* chorismate mutase, MECP synthase, and PurM N-terminal domain) form homo- trimers or dimers and the β-sheets of the trimer or dimer form a barrel-shaped interface. Their active sites are placed in the clefts between adjacent monomers (Figure 3.3a) and thus are constrained to the edges of the α/β sandwich for each monomer. Although protein domains of the gp5 domain A family do not bind substrates or ligands, they do interact with each other, participating in formation of homo-hexamers or pentamers stabilized partially by electrostatic and hydrophobic interactions between α-helices α1 and α2 of adjacent monomers (Figure 3.1f).

While homologous protein domains usually have similar active site locations, we found an unusual exception. As we mentioned in the family description (section II.10), the four thioredoxin-like domains of the AICAR Tfase part of the bifunctional purine biosynthesis enzyme ATIC are homologous to each other. The active site of AICAR Tfase is between the first thioredoxin-like domain of one monomer and the second thioredoxin-like domain of the other monomer (Figure 3.3b). The second thioredoxin-like domain houses active site residues at the loop regions near the N-terminal ends of the α-helices, similarly to most other thioredoxin-like domains (type i location); while the first thioredoxin-like domain has active site residues in the loop regions near the C-terminal ends of the α-helices, with two catalytic residues located at the loop between β-strands 3 and 4 (Wolan, Greasley et al. 2002), which is opposite to that of the second domain (Figure 3.3b). Thus, our analysis reveals a rare example of homologous protein domains possessing different active site locations.

3.5.2 Comparison to other structure classifications

Different structure classifications use different criteria and methods. The protein domains that we unified in the thioredoxin-like fold group are categorized differently in three major structure classifications CATH (Orengo, Michie et al. 1997; Pearl, Lee et al. 2000; Orengo, Bray et al. 2002), SCOP (Murzin, Brenner et al. 1995; Lo Conte, Ailey et al. 2000; Lo Conte, Brenner et al. 2002), and Dali Domain Dictionary (Holm and Sander 1996; Holm and Sander 1998; Dietmann and Holm 2001).

In CATH (version 2.5), some of these thioredoxin-like fold protein domains are not classified at all, such as the 2C-methyl-D-erythritol-2,4-cyclodiphosphate synthase, the AICAR transfermylase domain of bifunctional purine biosynthesis enzyme ATIC, and the capsid gp5 protein domain A; the others are placed into five fold groups (CATH "topology" level). Three of the fold groups correspond to three different circular permutations, and protein domains of type I circular permutation are divided into two fold groups. CATH assigns the small domain of RNA 3'-terminal phosphate cyclase (RTPC) to a different fold group than the thioredoxin proteins, although they both have the same type of circular permutation. In fact, CATH classifies them into two different architecture types (a higher level in the classification hierarchy than fold groups): a 2-layer sandwich and a 3-layer sandwich. The other fold groups are also categorized as 2- or 3-layer sandwich architecture types. CATH groups our thioredoxin-like protein domains into nine homologous superfamilies, which is basically consistent with our evolutionary family classification except for one protein. We assigned the C-terminal domain of phenol hydroxylase (1fohA) to be in the same evolutionary family as the classical thioredoxin-like proteins, while CATH assigns it into a separate superfamily by itself.

SCOP (version 1.65) classifies the thioredoxin-like fold domains into five different fold groups (SCOP "fold" level) corresponding to four different circular permutations and one separate fold group for the entire capsid protein gp5. SCOP does not break gp5 into domains; instead, it assigns the entire gp5 protein to a separate fold group and describes it as an unusual fold. The small domain of RTPC is assigned to the same fold group as the thioredoxin proteins in SCOP. SCOP fold groups are placed into two different structural classes: α/β and $\alpha+\beta$. At the evolutionary family level, our classification is consistent with SCOP superfamily classification.

Dali Domain Dictionary (DaliDD, version 3.1 beta) classifies the thioredoxin-like fold protein domains into seven fold groups (DaliDD "globular folding topology" level). In this classification, there are protein domains of the same circular permutation assigned to different fold groups, such as the N-terminal domain (1a8l_1) and the C-terminal domain (1a8l_2) of an archaeon PDI; there are also protein domains of different circular permutations assigned to the same fold group, such as the C-terminal domain of two-domain cytidine deaminase (1aln_2) and the cell division protein FtsZ (1fsz). DaliDD splits the thioredoxin-like protein domains into many more evolutionary families than we do. For example, the protein domains in one of our evolutionary family, the thioredoxin family, are placed into seven functional families (the highest hierarchy indicating evolutionary relationships in DaliDD). Nevertheless, DaliDD classifies the C-terminal domain of phenol hydroxylase (1fohA) into the same functional family as glutathione peroxidase (1gp1A), one of the classical thioredoxin-like proteins.

From the above comparisons, we perceive that the discrepancies between different structure classifications of these thioredoxin-like fold proteins mainly arise from the problems of the definition of the thioredoxin-like fold (2- or 3-layer sandwich) and the treatment of different circular permutations. By defining the structural core of the thioredoxin-like fold and considering different circular permutations within the same fold group, we resolve the discrepancies between the structure classifications. Grouping all these structurally similar thioredoxin-like proteins together enables us to study their evolutionary relationships and functional properties, which should be helpful for structure-functional predictions of uncharacterized thioredoxin-like fold proteins.

3.5.3 Structural analogs

During our structure search, we encountered a number of protein domains with the thioredoxin structural motif that we did not include in our thioredoxin fold group. Although these proteins were found by automatic searches for the thioredoxin fold, since they contain all the required secondary structure elements and interactions between them, we believe that they belong to fold groups other than the thioredoxin-like fold group based on the reasoning below.

Homology relationship determined structural core selection

Peptide methionine sulfoxide reductase (PMSR) contains two overlapping structural motifs: the thioredoxin-like motif and the ferredoxin-like motif. Figure 3.4c shows a typical ferredoxin-like fold protein. It is an α/β sandwich with the βαββαβ secondary structure pattern. The four β-strands ordering 2314 form an anti-parallel β-sheet with the two α-helices on one side. From Figure 3.4a, we can see that if we treat α-helix αA and β-strand βB as insertions, PMSR adopts a thioredoxin-like fold of type III circular permutation. On the other hand, if we treat α-helix α1 and β-strand β2 as insertions, the protein adopts a ferredoxin-like fold (Figure 3.4b). α1, β2 and αA, βB are placed on different sides of the central β-sheet and thus occupy similar positions in relation to the structure core. β2 and βB have approximately the same length (Figure 3.4a & Figure 3.4b). Thus if we try to base our decision about the fold solely on the structural properties of this molecule, both structural motifs (thioredoxin-like and ferredoxin-like) appear reasonable and we are unable to choose one of them. Sequence analysis, however, shows that PMSR is homologous to the ferredoxin-like fold protein that is shown in Figure 3.4c, the fourth metal-binding domain of Menkes copper-transporting ATPase (L.N. Kinch and N.V. Grishin, unpublished), that is missing α-helix α1 and β-strand β2 and hence missing the thioredoxin-like motif. Therefore, the ferredoxin-like motif is the essential one for PMSR, and PMSR most likely obtained the thioredoxin-like motif later in the process of evolution by insertions of α-helix α1 and β-strand β2. In spite of the thioredoxin-like motif in PMSR structure, it should be classified in a ferredoxin-like fold. Using similar reasoning we ruled out the following proteins with the thioredoxin motif: C-terminal domain of glyceraldehyde-3-phosphate dehydrogenase (1a7kA), transcription factor sc-mtTFB (1i4wA), and histidyl-tRNA synthetase (1adjA).

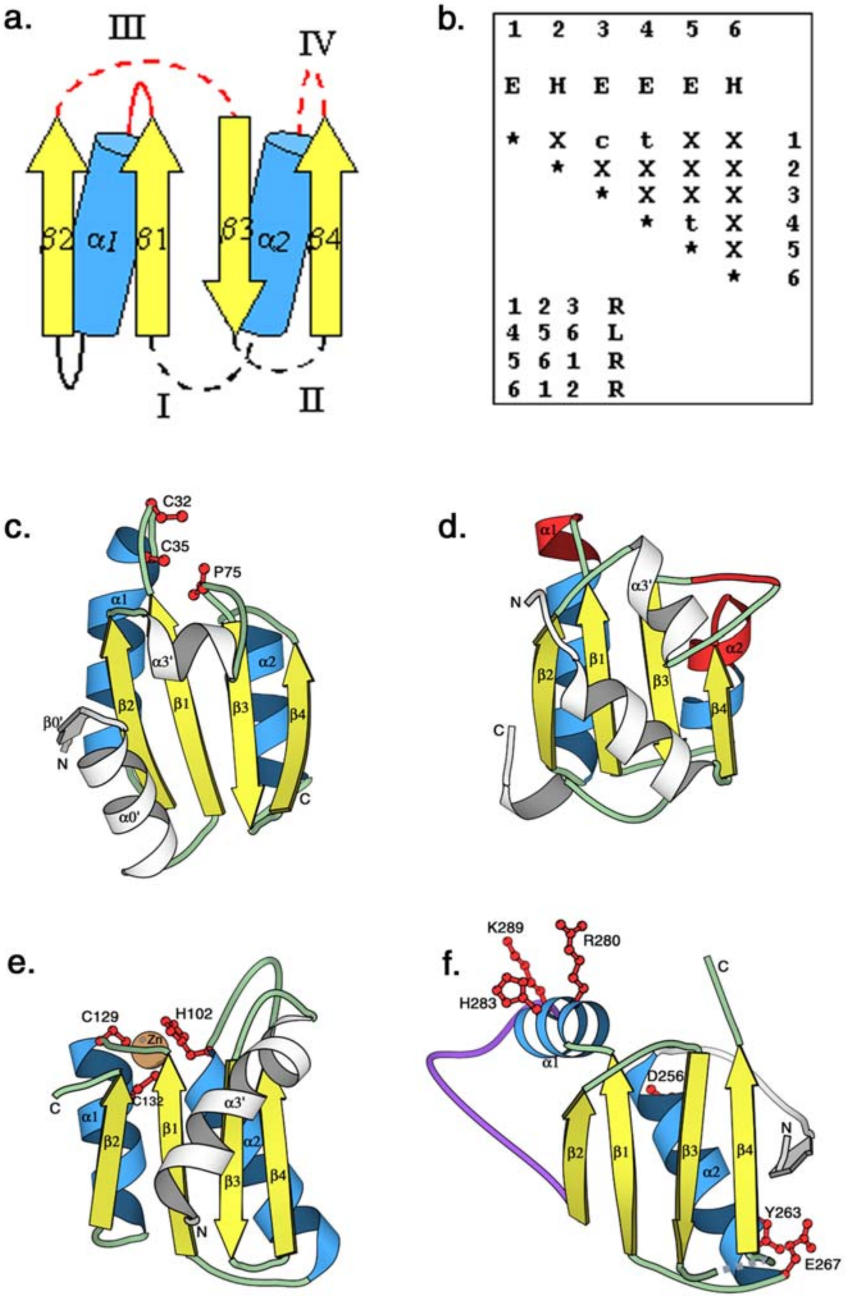
Structural importance determined structural core selection

The C-terminal domain of subunit A of the archaeon formylmethanofuran: tetrahydromethanopterin formyltransferase (Ftr) contains a thioredoxin-like motif of the β2β3β4α2β1α1 circular permutation if α-helix αA and β-strands βB and βB' are treated as insertions (Figure 3.4d). If we include α-helix αA and β-strands βB and βB' and treat α-helix α1 and β-strand β2 as insertions, this domain adopts a ferredoxin-like fold (Figure 3.4e). Weather or not to assign this protein domain into the thioredoxin-like fold group depends on which group of secondary structures we treat as insertions: αA, βB and βB', or α1 and β2. Comparisons between αA, βB, βB' and α1, β2 shows that αA, βB, βB' are seemingly more important (i.e. core) secondary structure elements than α1 and β2. αA is a 16 residues long α-helix that extensively interacts with the 18 residues long central α-helix α2; while α1 is only a 6 residues long, one and half-turn α-helix that interacts with the central α-helix α2 through just a few residues. The average length of the three central β-strands β1, β3 and β4 is about 9 residues long. If we consider βB and βB' as one β-strand interrupted by a loop, it is 7 residues long and forms 5 hydrogen bonds with one of the central β-strands β4; while β2 is 4 residues long and forms only 2 hydrogen bonds with β1. Therefore, αA, βB and βB' are more important secondary structure elements than α1 and β2, and thus should not be treated as insertions. Hence, the structural core of the protein domain is not formed by the thioredoxin-like secondary structure elements. In addition, at the crossover of the loops connecting β1-α1 (L1) and β2-β3 (L2), L1 is below L2 in a typical thioredoxin-like fold in the orientation shown in Figure 3.1d, while L1 is above L2 in Ftr in the same orientation shown in Figure 3.4d. As a result, although Ftr contains the thioredoxin-like motif, it is not a thioredoxin-like fold protein, but a ferredoxin-like fold protein (Figure 3.4e). The reasoning for ruling out the N-terminal domain of subunit A of Ftr (1ftrA), the catalytic domain of type 1 cytotoxic necrotizing factor (1hq0A), and replication terminator protein (1ecrA) is similar.

3.6 CONCLUSIONS

A hierarchical structure classification of thioredoxin-like fold proteins has been carried out. We define the thioredoxin-like fold and identify 723 protein domains as thioredoxin-like fold. These domains are grouped into eleven evolutionary families. A structure-based multiple sequence alignment of 90 representative thioredoxin motif-containing proteins is manually constructed. Analysis of the secondary structure connectivity identifies four types of circular permutations and a potential functional/packing unit. Analysis of active site locations reveals two major functional sites for the thioredoxin motif-containing proteins and one rare example of homologous protein domains possessing different functional sites. Comparison to existing structure classifications shows that our thioredoxin-like fold group is broader and more inclusive.

Figure 3.1 Thioredoxin-like Fold and Its Observed Circular Permutations



Thioredoxin-like fold and its observed circular permutations. (a) The topological diagram of the thioredoxin-like fold. α-helices and β-strands are shown as blue cylinders and yellow arrows, respectively, the lines connecting different secondary structures represent loop regions between them. Dotted loops indicate the termini positions of the four types of circular permutations that we observed. No termini were observed at solid loop locations. Loops shown in red indicate the type i active site location. (b) The query matrix of the thioredoxin-like fold of type I circular permutation. Secondary structures are consecutively numbered in Arabic numbers. Upper case letters E (β-strand) and H (α-helix) indicate the type of secondary structure. Lower case letters c and t indicate parallel and anti-parallel hydrogen-bonding interactions between secondary structures, respectively. Upper case letter X indicates that no interactions were considered. Upper case letters R and L indicate right-handed and left-handed chirality in a triplet of secondary structures, respectively. Ribbon diagrams of (c) human thioredoxin (1ert (Weichsel, Gasdaska et al. 1996)), a representative of type I circular permutation, (d) yeast ribosomal protein L30 (1cn8A (Mao, White et al. 1999)), a representative of type II circular permutation, (e) *E. coli* cytidine deaminase (1aln_1 (Xiang, Short et al. 1996)), a representative of type III circular permutation, and (f) bacteriophage HK97 capsid protein gp5 (1ohg (Helgstrand, Wikoff et al. 2003), previous PDB ID: 1fh6), a representative of type IV circular permutation, were produced using the program MOLSCRIPT (Kraulis 1991). Corresponding secondary structure elements are colored and named as in diagram (a). Elements corresponding to inserted domains are shown in white. The long insertion in capsid protein gp5 is shown in purple in (f). In (c), (e), and (f), active site residues are depicted in red ball-and-stick representation. In (d), active site residues interacting with RNA are shown in red. The orange sphere in (e) shows a zinc ion.

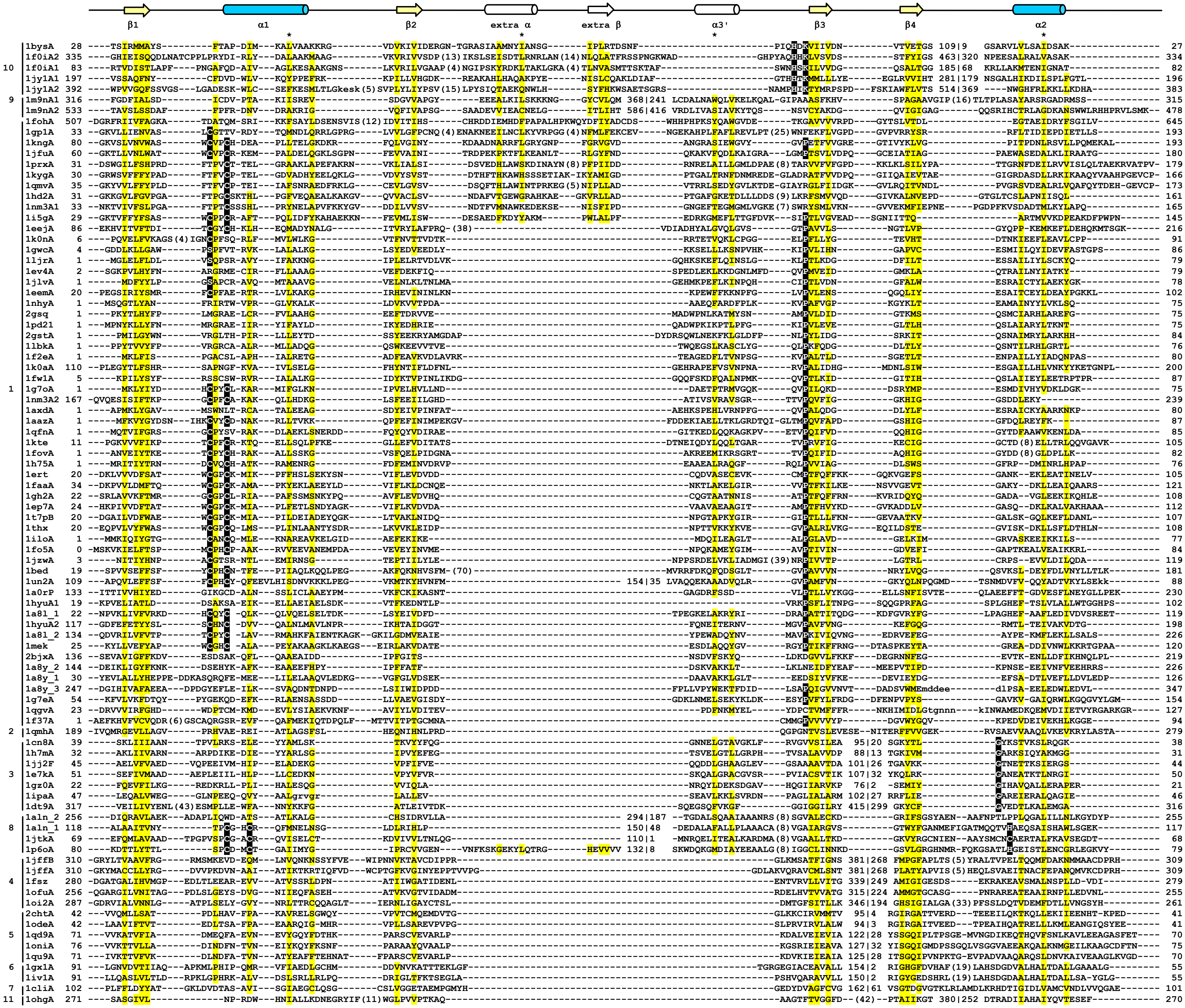
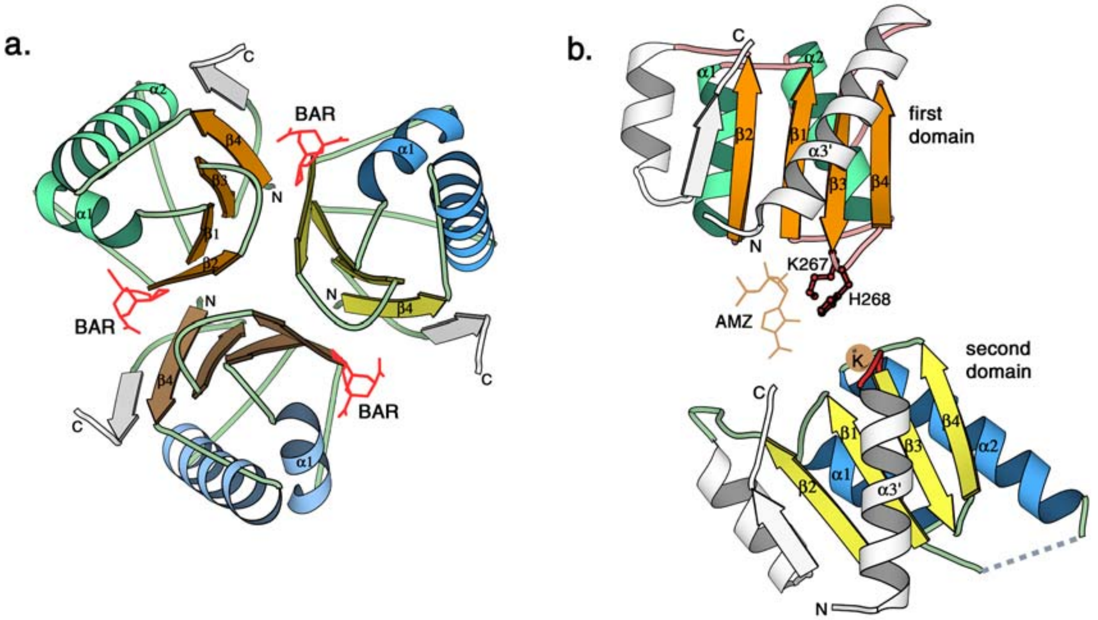

β1
α1
β2
extra α
extra β
α3'
β3
β4
α2
10
9
1
2
3
8
4
5
6
7
11
1bysA
1f0iA2
1f0iA1
1jy1A1
1jy1A2
1m9nA1
1m9nA2
1fohA
1gp1A
1kngA
1jfuA
1prxA
1kygA
1qmvA
1hd2A
1nm3A1
1i5gA
1eejA
1k0nA
1gwcA
1ljrA
1ev4A
1jlvA
1eemA
1nhyA
2gsq
1pd21
2gstA
1lbkA
1f2eA
1k0aA
1fw1A
1g7oA
1nm3A2
1axdA
1aazA
1qfnA
1kte
1fovA
1h75A
1ert
1faaA
1gh2A
1ep7A
1t7pB
1thx
1iloA
1fo5A
1jzwA
1bed
1un2A
1a0rP
1hyuA1
1a8l_1
1hyuA2
1a8l_2
1mek
2bjxA
1a8y_2
1a8y_1
1a8y_3
1g7eA
1qgvA
1f37A
1qmhA
1cn8A
1h7mA
1jj2F
1e7kA
1gz0A
1ipaA
1dt9A
1aln_2
1aln_1
1jtkA
1p6oA
1jffB
1jffA
1fsz
1ofuA
1oi2A
2chtA
1odeA
1qd9A
1oniA
1qu9A
1gx1A
1iv1A
1cliA
1ohgA

Figure 3.2 Structure-based multiple sequence alignment of representative thioredoxin-like protein domains

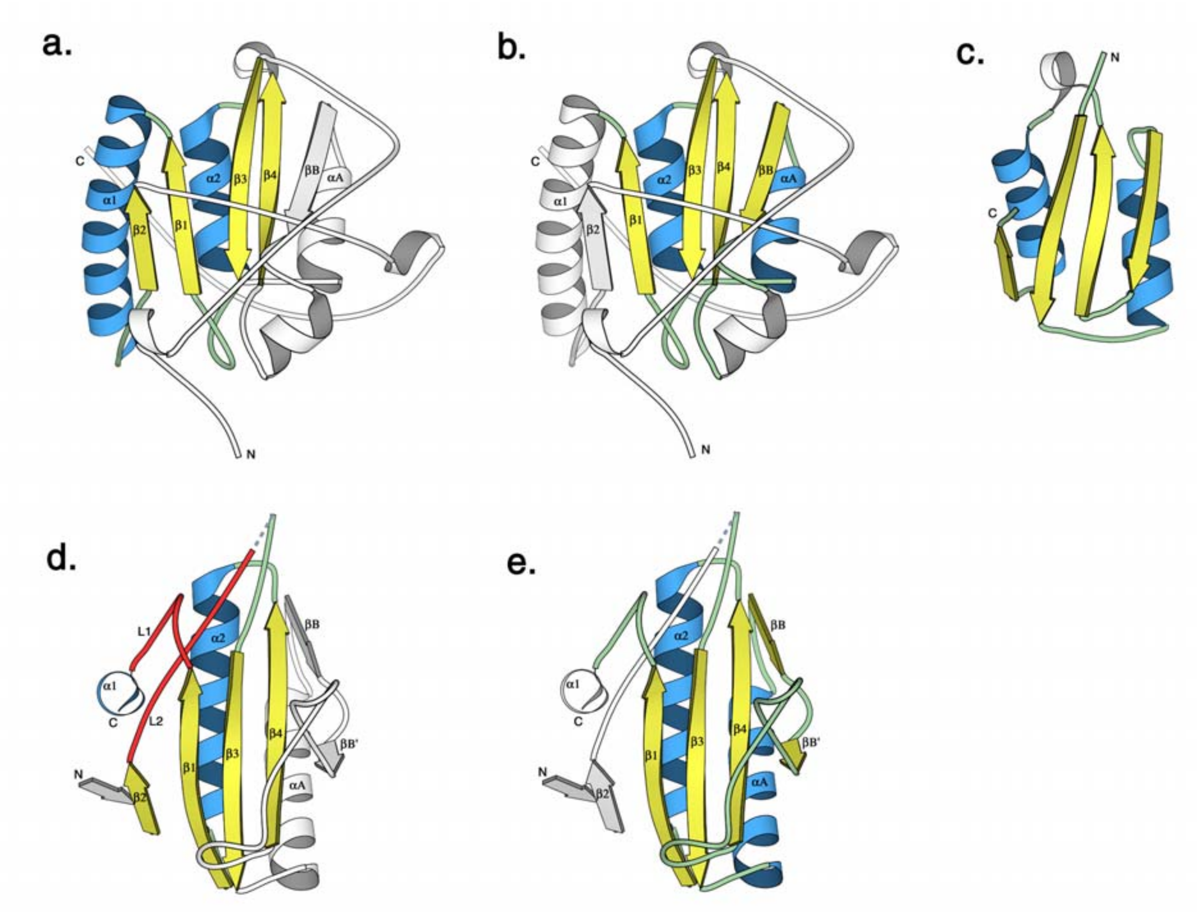
(Please see the previous page for the Figure) Each sequence is labeled by its PDB identifier followed by an optional chain identifier at the 5th position and an optional domain identifier for duplicated domains at the 6th position. Sequences are grouped according to 11 evolutionary families. The first and the last residue numbers are indicated for each sequence. Sequences of type II, III and IV circular permutations are rearranged to align their corresponding secondary structure elements with the type I circular permutation. The termini in these proteins are separated by a "|" and the residue numbers around the permuted region are shown in red. Long insertions in loop regions are omitted with the number of missing residues in parentheses. Sequences in lower case represent disordered regions in structures. Sequences in italics differ in secondary structure from the consensus secondary structure of the alignment. Uncharged residues at mainly hydrophobic positions are highlighted in yellow and magenta asterisks mark the hydrophobic positions that were used to aid alignment of α-helices. Conserved residues within each family are highlighted in black. The diagram of secondary structures (α-helices as cylinders and β-strands as arrows) is shown above the alignment. Representative protein sequences of each evolutionary family are included in the alignment. They are as follows. 1. phenol hydroxylase C-terminal domain (1fohA), glutathione peroxidase (1gp1A), cytochrome c maturation oxidoreductase CcmG (1kngA), soluble domain of membrane-anchored thioredoxin-like protein TlpA (1jfuA), peroxiredoxins (1prxA, 1qmvA, 1hd2A, 1nm3A1), alkyl hydroperoxide reductase AhpC (1kygA), tryparedoxin (1i5gA), disulfide bond isomerase DsbC C-terminal domain (1eejA), chloride intracellular channel 1 clic1 (1k0nA), glutathione S-transferases (1gwcA, 1ljrA, 1ev4A, 1jlvA, 1eemA, 2gsq, 1pd21, 2gstA, 1lbkA, 1f2eA, 1fw1A, 1axdA), GST-like domain of elongation factor 1-gamma (1nhyA), nitrogen regulation fragment of yeast prion protein ure2p (1k0aA), glutaredoxins (1g7oA, 1nm3A2, 1aazA, 1qfnA, 1kte, 1fovA), NrdH-redoxin (1h75A), thioredoxins (1ert, 1faaA, 1gh2A, 1ep7A, 1t7pB, 1thx, 1iloA), thioredoxin/glutaredoxin-like protein MJ0307 (1fo5A), arsenate reductase ArsC (1jzwA), disulphide bond oxidases DsbA (1bed, 1un2A), phosducin (1a0rP), Alkyl hydroperoxide reductase subunit F AhpF N-terminal domain (1hyuA1, 1hyuA2), protein disulfide isomerases (1a8l_1, 1a8l_2, 1mek, 2bjxA), calsequestrin (1a8y_2, 1a8y_1, 1a8y_3), endoplasmic reticulum protein ERP29 N-terminal domain (1g7eA), spliceosomal protein U5-15Kd (1qgvA), thioredoxin-like 2Fe-2S ferredoxin (1f37A); 2. small domains of the RNA 3'-terminal phosphate cyclase (1qmhA); 3. eukaryotic ribosomal protein L30e (1cn8A, 1h7mA), ribosomal protein L7ae (1jj2F), spliceosomal 15.5kd protein (1e7kA), RNA 2'O-methyltransferases N-terminal domain (1gz0A, 1ipaA), eukaryotic peptide chain release factor subunit 1 ERF1 C-terminal domain (1dt9A); 4. tubulin β-subunit (1jffB), tubulin α-subunit (1jffA), cell-division proteins FtsZ (1fsz, 1ofuA), dihydroxyacetone kinase subunit K (1oi2A); 5. chorismate mutases (2chtA, 1odeA), purine regulatory protein YabJ (1qd9A), translational Inhibitor Protein P14.5 (1oniA), hypothetical protein YjgF (1qu9A); 6. 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthases (1gx1A, 1iv1A); 7. aminoimidazole ribonucleotide synthetase N-terminal domain (1cliA); 8. two-domain CDA (1aln_1, 1aln_2), one-domain cytidine deaminase (1jtkA), cytosine deaminase (1p6oA); 9. AICAR transformylase domain of bifunctional purine biosynthesis enzyme ATIC (1m9nA1, 1m9nA2); 10. nuclease Nuc (1bysA), phospholipase D (1f0iA1, 1f0iA2), tyrosyl-DNA phosphodiesterase TDP1 (1jy1A1, 1jy1A2); 11. domain A of capsid protein gp5 (1ohgA).

Figure 3.3 Active site locations



Active site locations. (a) Ribbon diagram of bacillus chorismate mutase (2cht (Chook, Ke et al. 1993)) with its substrate analogs shows the type ii active site location. The α-helices and β-strands are numbered as in Figure 3.1a, but are colored differently in different domains with inserted elements in white, and substrate analog BAR in red. The three domains are viewed along their three-fold axis. One BAR molecule is located at each of the three clefts between two adjacent domains. (b) Ribbon diagram of two thioredoxin-like domains in the AICAR Tfase part of bifunctional purine biosynthesis enzyme ATIC (1m9n (Wolan, Greasley et al. 2002)) illustrates an unusual active site location. The second domain of one monomer is colored as in Figure 3.1a, while the first domain of another monomer is shown in a different color scheme. The other two thioredoxin-like domains of the AICAR Tfase part are omitted for clarity. The substrate AMZ is shown in brown and marks the active site. Two catalytic residues from the first domain are shown as ball-and-stick in red. A potassium ion represented by an orange sphere binds to the loop (shown in red) between α3' and β4 of the second domain. Both ribbon diagrams were generated using the program MOLSCRIPT (Kraulis 1991).

Figure 3.4 Structure Analogs



Structure analogs. Ribbon diagrams of (a, b) *E. coli* peptide methionine sulfoxide reductase (1ff3 (Tete-Favier, Cobessi et al. 2000)), (c) the fourth metal-binding domain of human Menkes copper-transporting ATPase (1aw0 (Gitschier, Moffat et al. 1998)), a ferredoxin-like fold protein, and (d, e) archaeon formylmethanofuran:tetrahydromethanopterin formyltransferase (1ftr (Ermler, Merckel et al. 1997)). Protein domains in (a) and (b) are the same, however, in (a), the elements of the thioredoxin-like motif are colored in yellow and blue; in (b), the elements of the ferredoxin-like motif are colored in yellow and blue. Similarly, we colored the 1ftr domain in (d) and (e). In (d), the loops L1 and L2 are shown in red. All ribbon diagrams were generated using the program MOLSCRIPT (Kraulis 1991).

Table 3.1 Summary of Thioredoxin-like Domain Families

Family	Circular permutation type	Active site location type	Representatives in the alignment
1. Thioredoxin	I	i	1fohA, 1gp1A, 1kngA, 1jfuA, 1prxA, 1kygA, 1qmvA, 1hd2A, 1nm3A, 1i5gA, 1eejA, 1k0nA, 1gwcA, 1ljrA, 1ev4A, 1jlvA, 1eemA, 1nhyA, 2gsq, 1pd21, 2gstA, 1lbkA, 1f2eA, 1k0aA, 1fw1A, 1g7oA, 1axdA, 1aazA, 1qfnA, 1kte, 1fovA, 1h75A, 1ert, 1faaA, 1gh2A, 1ep7A, 1t7pB, 1thx, 1iloA, 1fo5A, 1jzwA, 1bed, 1un2A*, 1a0rP, 1hyuA, 1a8l, 1mek, 2bjxA, 1a8y, 1g7eA, 1qgvA, 1f37A
2. RTPC small domain	I	Unknown	1qmhA
3. Ribosomal protein L30e	II	i	1cn8A, 1h7mA, 1jj2F, 1e7kA, 1az0A, 1ipaA, 1dt9A
4. Tubulin C-terminal domain	II	i	1jffB, 1jffA, 1fsz, 1ofuA, 1oi2A
5. Bacillus chorismate mutase	II	ii (trimer)	2chtA, 1odeA, 1qd9A, 1oniA, 1qu9A
6. MECP synthase	II	ii (trimer)	1gx1A, 1iv1A
7. PurM	II	ii (dimer)	1cliA
8. Cytidine deaminase	III	i	1aln, 1jtkA, 1p6oA
9. AICAR Tfase domain of ATIC	III	Unusual	1m9nA
10. Phospholipase D	IV	i	1bysA, 1f0iA, 1jy1A
11. Gp5 domain A	IV	ii (hexamer or pentamer)	1ohgA**

* Previous PDB identifier: 1dyv

** Previous PDB identifier: 1fh6

CHAPTER 4:
PCOAT: A Tool For Protein Positional Correlation Analysis

4.1 INTRODUCTION

4.1.1 Background

Positional correlation or covariation refers to the phenomenon that mutations at one position of a protein influence the mutations at other positions of the protein during evolution. Correlation between positions may arise for structural or functional reasons, such as stabilizing local contact (Mateu and Fersht 1999) or affecting protein functions through networks of interactions (Suel, Lockless et al. 2003). Different methods have been developed to detect and evaluate positional correlations in a multiple sequence alignment, including approaches based on mutual information (Crowder, Holton et al. 2001), chi-square test (Larson, Di Nardo et al. 2000), and correlation coefficient (Saraf, Moore et al. 2003). Each method has its advantages and limitations (Pollock and Taylor 1997). In addition, distinguishing structurally or functionally important correlations from background correlations caused by phylogeny or stochastic events remains difficult (Atchley, Wollenberg et al. 2000).

4.1.2 Objective

Aiming at the problems of the existing methods, we have developed a program (Positional Correlation Analysis Tool) that performs positional correlation analysis comprehensively and systematically. We have implemented different statistical significance estimation methods to identify correlated position pairs, amino acid pairs and networks of correlated positions in an input alignment, and utilized multiple sequence weighting and sampling methods to eliminate the background correlations. Our program should be useful and convenient for researchers to identify candidate residues for structurally or functionally important interactions in their protein families.

4.2 ALGORITHMS

For an input multiple sequence alignment, PCOAT performs the positional correlation analysis in four steps. First, the effective count of every amino acid pair at each position pair is estimated. Second, correlation scores of every position pair and amino acid pair are determined with corresponding statistical significances and the pairs that are significantly correlated are identified. Next, individual positions that are highly correlated with multiple other positions are detected, and an optional fourth step identifies the networks of highly correlated positions.

4.2.1 Estimation of effective counts

In order to eliminate background correlations (i.e. help to remove phylogenetic artifact) and correct for redundant sequences in the input alignment, we implemented three weighting methods to estimate the effective count of every amino acid pair at each position pair: unweighted count, Henikoff weighting (HW) count (Henikoff and Henikoff 1994), and altered position-specific independent count (PSIC) (Sunyaev, Eisenhaber et al. 1999; Pei and Grishin 2001). Both HW and PSIC methods have been modified to calculate the weight for a position pair instead of for a single position. The estimated effective counts are stored in contingency tables. Invariant positions and gapped positions are removed to eliminate potential false positives.

4.2.2 Identification of statistically significantly correlated position pairs and amino acid pairs

To identify significantly correlated position pairs and amino acid pairs, we have implemented two statistical tests: Pearson's chi-square test of independence and likelihood ratio test. Both tests have been proved useful for positional correlation analysis (Larson, Di Nardo et al. 2000; Crowder, Holton et al. 2001).

Pearson's Chi-square test

Our null hypothesis is that the amino acid substitutions at any two positions are independent of each other. Based on this hypothesis, we have $f_{ij}^{ab} = f_i^a * f_j^b$, where f_{ij}^{ab} is the expected frequency of amino acid *a* occurring at position *i* and amino acid *b* occurring at position *j* in the same sequences, f_i^a is the observed frequency of amino acid *a* at position *i*, and f_j^b is the observed frequency of amino acid *b* at position *j*. Thus, the expected count of amino acid pair *a* and *b* at position pair *i* and *j* is $e_{ij}^{ab} = f_{ij}^{ab} * T_{ij}$, where T_{ij} is the total effective number of sequences at position pair *i* and *j*. Pearson's chi-square statistic χ^2 is calculated as $\chi_{ij}^2 = \sum_{a,b=1..20} \frac{\left(n_{ij}^{ab} - e_{ij}^{ab}\right)^2}{e_{ij}^{ab}}$ for position pair *i* and *j*, and $\chi_{ij,pq}^2 = \sum_{\substack{a=p,\bar{p} \\ b=q,\bar{q}}} \frac{\left(n_{ij}^{ab} - e_{ij}^{ab}\right)^2}{e_{ij}^{ab}}$ for amino acid pair *p* and *q*, where n_{ij}^{ab} is the observed effective count of amino acid pair *a* and *b* at position pair *i* and *j*, and *a* and *b* take the value of residue type *p* and $\bar{p}$ (all residue types but *p*), and *q* and $\bar{q}$ (all residue types but *q*), respectively. χ^2 is utilized to measure the fit between the observed counts and the expected counts, and obeys the chi-square distribution when the null hypothesis (no correlation) is true. We calculate the chi-square probability of χ^2 for amino acid pairs, and the Z-score of χ^2 for position pairs.

Sampling the alignment by vertical shuffling

The mean and standard deviation parameters of a random χ^2 distribution that are required to calculate the Z-scores can be the theoretical mean and standard deviation of the chi-square distribution, or can be obtained from sampling the alignment by vertical shuffling. Calculating Z-scores using parameters obtained by sampling helps eliminate amino acid composition bias, and thus increases the prediction accuracy.

Likelihood ratio test

Likelihood ratio statistic G^2 is calculated as $G_{ij}^2 = 2\sum_{a,b=1..20} n_{ij}^{ab} \log \frac{n_{ij}^{ab}}{e_{ij}^{ab}}$ for position pair i and j, and $G_{ij,pq}^2 = 2\sum_{\substack{a=p,\bar{p} \\ b=q,\bar{q}}} n_{ij}^{ab} \log \frac{n_{ij}^{ab}}{e_{ij}^{ab}}$ for amino acid pair p and q. G^2 also follows the chi-square distribution under the null hypothesis. We calculate chi-square probabilities of G^2 for amino acid pairs and Z-scores of G^2 for position pairs the same way as for χ^2.

4.2.3 Identification of highly correlated positions

To identify individual positions that are highly correlated with other positions in the input alignment, we calculate the χ^2 and G^2 of each position i as $\chi_i^2 = \sum_{\substack{j=1..N \\ j\neq i}} \chi_{ij}^2$, $G_i^2 = \sum_{\substack{j=1..N \\ j\neq i}} G_{ij}^2$ based on the addition theorem of chi-square distribution, where N is the total number of positions in the alignment. The Z-scores of χ_i^2 and G_i^2 of each position are calculated and ranked by their statistical significance.

4.2.4 Identification of networks of correlated positions

To identify potential networks of correlated positions, we implement two clustering methods. Groups of inter-correlated positions can be detected by single-linkage or complete-linkage clustering when the degrees of correlations between them are higher than a user-definable significance threshold. When using single-linkage clustering, the data points are loosely connected and the average distance between data points could be long. When using complete-linkage clustering, the data points are more compactly connected with each other and the average distance between data points is shorter compared to single-linkage clustering. User could select different options according to their needs.

4.3 APPLICATIONS AND COMPARISON TO OTHER METHODS

Comparison to other correlation analysis programs, DEPENDENCY (Tillier and Lui 2003) and CRASP (Afonnikov, Oshchepkov et al. 2001), shows that PCOAT is the only correlation analysis program that identifies correlated position pairs as well as correlated amino acid pairs, highly correlated positions, and networks of correlated positions. In addition, PCOAT runs faster (1.5-5 times depending on the family size) and is capable to analyze alignments with large number of sequences (more than 10,000), while other programs cannot. We have applied PCOAT to a number of Pfam (Bateman, Coin et al. 2004) alignments. Analysis of the C2H2 zinc finger family alignment (28,239 sequences, April 2004) using PCOAT identified positions 52 and 57 as the most correlated position pair, and residues Arg52 and Asp57 at these two positions as the most correlated amino acid pair (Figure 4.1). It has been shown that positions 52 and 57 are important substrate specificity determinants for zinc finger binding to dsDNA (Iuchi 2001). Arg52 and Asp57, binding to G and C or A of the substrate, respectively, are the dominant amino acids at positions 52 and 57 in the triple-C2H2 class of zinc finger proteins (Iuchi 2001). Both DEPENDENCY (Tillier and Lui 2003) and CRASP (Afonnikov, Oshchepkov et al. 2001) cannot analyze this C2H2 alignment because of the large number of sequences. Working on a reduced-size (10,000)

C2H2 alignment, DEPENDENCY detected positions 52 and 57 as one of the four significantly correlated position pairs but took 4.75 times the running time of PCOAT (Table 4.1). Analysis of the ACT domain alignment (1380 sequences, April 2004) using PCOAT identified a network of nine correlated positions forming a surface patch on the β-strands of the domain (Figure 4.2). Structure analysis of the phenylalanine hydroxylase shows that the ACT domain interacts with the catalytic domain through the β-strands (Kobe, Jennings et al. 1999). The correlation network identified by PCOAT possibly plays a role in this interaction.

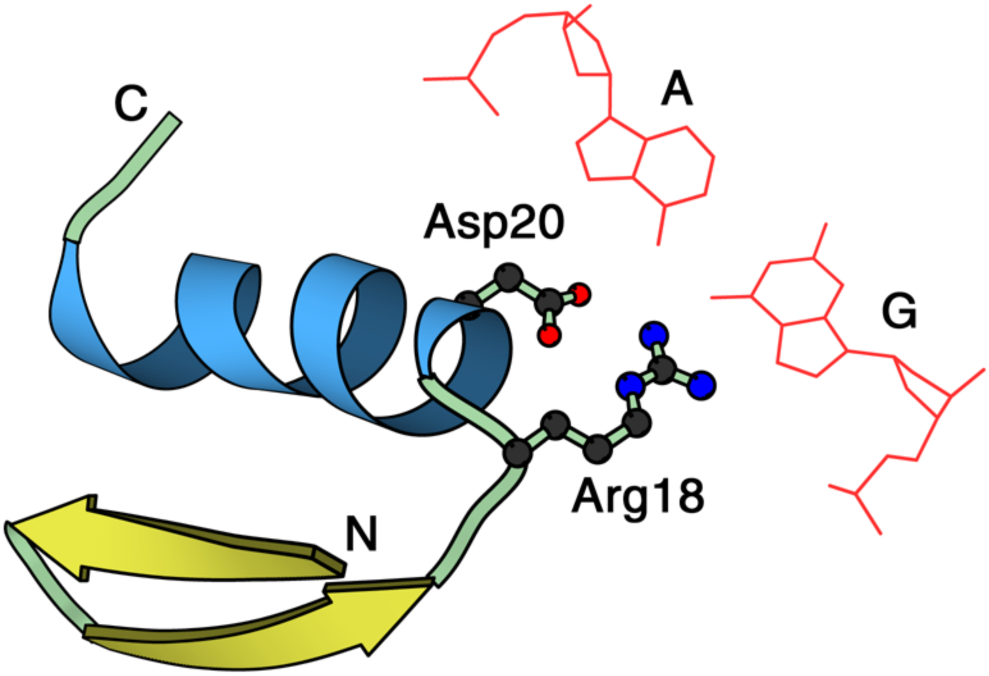
4.4 PROGRAM AVAILABILITY

The source code and executable files for different operating systems of PCOAT are available for download at ftp://iole.swmed.edu/pub/PCOAT/. This ftp site also contains a detailed description of the program and the complete results of PCOAT analysis on C2H2 alignment and ACT domain alignment.

4.5 CONCLUSIONS

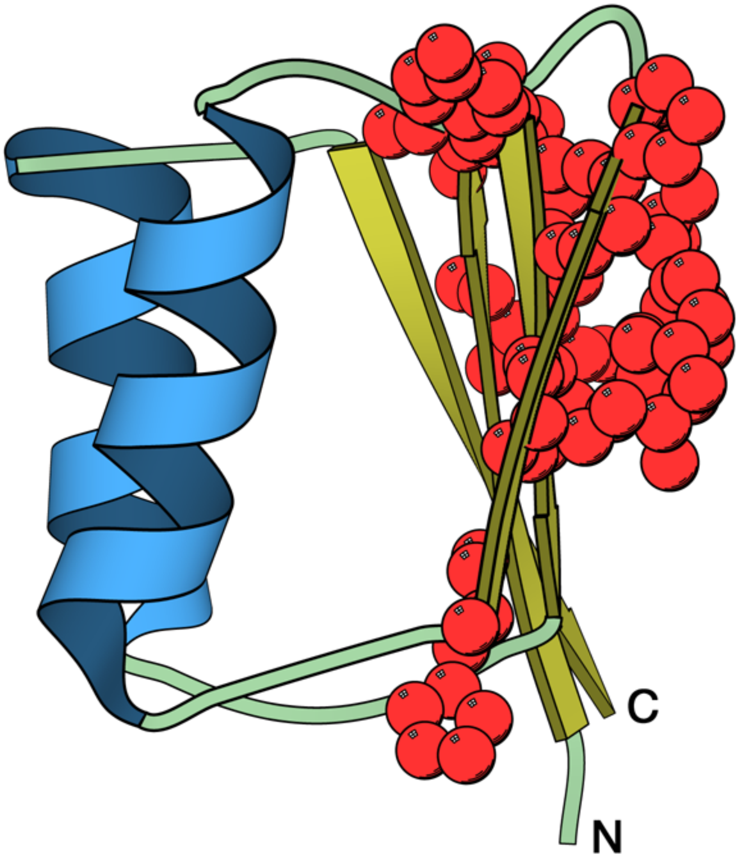
A computer program, PCOAT (Positional Correlation Analysis Tool), has been developed to perform positional correlation analysis for protein multiple sequence alignment in order to identify structurally or functionally important interactions between positions in a protein family. Different statistical methods have been implemented to detect highly correlated position pairs, amino acid pairs, individual positions, and networks of correlated positions, and developed multiple sequence weighting and sampling methods to eliminate background correlations caused by phylogeny and stochastic events. This program runs relatively fast and is suitable for analyzing alignments containing large number of sequences. Applying PCOAT to protein families shows that the program is useful to identify structurally or functionally important residues.

Figure 4.1 Structure Diagram of C2H2 Zinc Finger Showing Correlated Pairs



Structure of a C2H2 zinc finger (1zaa_C:5-32) showing that residues Arg18 and Asp20 (Arg52 and Asp57 in the alignment) are in close local contact (2.7 Å) and interact with nucleotides G and A in the substrate, respectively.

Figure 4.2 Structure Diagram of ACT Domain Showing Correlated Network



The ACT domain (regulatory domain) of rat phenylalanine hydroxylase (1phz_A:33-111). Residues in the correlation network are shown as atom spheres.

Table 4.1 Comparison of Analysis Results Using PCOAT, DEPENDENCY and CRASP

	Top 3 significantly correlated	**PCOAT**	**DEPENDENCY**	**CRASP**
Zinc finger family alignment**	Positions	57 47 52	N/A	N/A
	Position pairs	52 and 57 53 and 60 52 and 58	4 and 63 52 and 57 16 and 63	47 and 26 52 and 72 64 and 72
	Amino acid pairs	R@52 and D@57* G@26 and K@46 E@16 and Q@63	N/A	N/A
	CPU time (s)	7.25	34.43	N/A
ACT domain alignment	Network or correlated positions	4 7 42 44 56 57 74 75 77	N/A	N/A
	CPU time (s)	34.53	39.31	N/A

* Means amino acid R at position 52 and amino acid D at position 57 are highly correlated.
** Zinc finger family alignment of reduced size (10,000 sequences).

CHAPTER 5:
Combining Sequence Profile With Predicted Secondary Structure For Structure Modeling And Homology Detection

5.1 INTRODUCTION

5.1.1 Background

Protein structures are important in terms of understanding the molecular mechanisms of their biological functions. Before the era of structural genomics, the protein structures available were biased in the structure and function space, stemming from experimental limitations and particular target selections by structural biologists (Xie and Bourne 2005). This bias results in the usage limitation of protein structure prediction by homology modeling method, since vast protein sequences have no nearby (i.e. of similar sequences) structures available. With the progress of structure genomics initiatives (Burley, Almo et al. 1999; Burley 2000), a set of landmark protein structures are being solved. These landmark protein structures are intended to fully map the protein structure and function space, rendering most newly discovered proteins within the homology modeling distance from a landmark structure (Lattman 2005). As a result, sequence-based homology modeling methods for protein structure prediction are of great practical importance.

In order to develop more powerful sequence similarity detection and alignment methods for structural and functional prediction purposes, we can add sequence-based predicted structural information on to the sequence profile information. Adding predicted secondary structure information to sequence profiles has been shown to help structure predictions by finding remote structure template (Kinch, Wrabl et al. 2003). Secondary structure prediction methods are basically mature now. PSIPRED (Jones 1999), one of the leading secondary structure prediction methods, is currently reported to have an average per residue accuracy (Q3) of ~78% (Eyrich, Marti-Renom et al. 2001; Bryson, McGuffin et al.

2005). The predicted secondary structures have been used in different studies or search methods (Ginalski, Pas et al. 2003; Tang, Xie et al. 2003; Ginalski, von Grotthuss et al. 2004; Teodorescu, Galor et al. 2004). Among them, HHsearch (Soding 2005) and Prof_ss (Chung and Yona 2004) are two most recently developed stand-alone sequence similarity search programs.

5.1.2 Objective

The purpose of our study is to improve sequence similarity search methods based on COMPASS (Sadreyev and Grishin 2003) (a profile-profile comparison method) by adding predicted secondary structure information for better distant similarity detection ability and alignment quality. There are different approaches to incorporate secondary structure information with sequence profiles. We compare their performance and find the best approach to use in our method.

5.2 ALGORITHM DEVELOPMENT

Our method is developed based on the theories of PSI-BLAST and COMPASS methods. The major improvement is to integrate predicted secondary structure information into the sequence profile. Compared to other methods, the strong point of our method is at the statistical significance estimation of the alignment scores. Four major steps are required in our algorithm to produce sequence alignments: (i) constructing substitution matrices of amino acid and secondary structure element; (ii) developing scoring function to score the combined sequence and secondary structure information between two matched positions; (iii) applying alignment algorithm to align two profiles and obtaining optimal alignment score; (iv) estimating statistical significance of the resulting optimal alignment.

5.2.1 Construction Of Substitution Matrices

Data used in substitution matrix constructions

The substitution matrices of amino acid and secondary structure element are constructed based on the sequence alignment data in the Blocks+ database (Henikoff and Henikoff 1991) and the predicted secondary structure data generated by PSIPRED (Jones 1999). Protein blocks are segments of ungapped multiple sequence alignments that correspond to the most highly conserved regions of the protein families. The Blocks+ database (Henikoff, Henikoff et al. 1999) version 13.0 consist of 11,853 blocks, of which 8656 blocks are taken from the Blocks database(Henikoff and Henikoff 1991) that are constructed automatically from the SWISS-PROT and TrEMBL sequences, and 3197 blocks are taken from the PRINTS database (version 31.0)(Attwood, Beck et al. 1994; Attwood, Bradley et al. 2003) that uses manual seeds followed by automatic methods. The secondary structures for each protein sequence in the Blocks+ database are predicted using PSIPRED, and the resulting predicted secondary structure elements, helix (H), extended strand (E) and unstructured coiled regions (C), are aligned with each other through the corresponding sequence alignment.

Clustering of sequences within blocks

In order to balance the information provided by multiple closely related sequences and single divergent sequences, we first need to weight the sequences in each block in the Blocks+ database. The weighting is done through clustering the sequences within each block by sequence identity using single-linkage clustering. Each resulting cluster is weighted as a single sequence (Henikoff and Henikoff 1992).

The BLOSUM series of amino acid substitution matrices shows that clustering at identity level 62% (BLOSUM62) gives best performance for sequence similarity detection. As the sequences in the Blocks+ database has expanded a great deal since the time when BLOSUM matrices were constructed, we screen for the optimal sequence identity level of clustering for the database we use by comparing values in our amino acid (AA) substitution matrix to BLOSUM62. An initial coarse screen of 45%, 62%, 80% and 90% followed by a refined screen around 80% find out that our AA substitution matrix of identity level 80% matches best with the values in BLOSUM62 (Figure 5.1). Therefore, we decide to construct all the substitution matrices at 80% sequence identity level.

Calculation of substitution matrices

Three types of substitution matrices are constructed based on these alignments: the 20x20 amino acid substitution matrix, the 3x3 secondary structure element substitution matrix and the 60x60 substitution matrix of combined amino acid and secondary structure symbols (e.g. If the predicted secondary structure element for amino acid Ala in a sequence is Helix, then Ala-helix is treated as one symbol. And there totally 60 combined symbols.). We will use the term symbol to refer to all three types of units in the substitution matrices, i.e. amino acid residue in the 20x20 substitution matrix, secondary structure element in the 3x3 substitution matrix, or the combined symbol (e.g. Ala-helix) in the 60x60 substitution matrix.

The substitution matrices are constructed the same way as the BLOSUM matrices (Henikoff and Henikoff 1992) with an extension to multiple dimensions (20x20, 3x3 and 60x60). Here we briefly describe the essential steps and formulae used in the construction (for more details, see Ref: Henikoff 1992). To construct the substitution matrices, the first step is to derive a count table that contains the observed counts of all possible pairs of symbols $f_{ij,\dim}$ (dim equals to 20, 3 and 60, respectively, for each type of substitution matrix) in all the clustered blocks. This is done column by column for each block of aligned symbols and then sums the counts up. Second is to derive the observed frequencies of symbol pairs occurring in an aligned position $q_{ij,\dim}$ from the observed count table using formula

Equation 5.1 $$q_{ij,\dim} = \frac{f_{ij,\dim}}{\sum_{i=1}^{\dim}\sum_{j=1}^{i} f_{ij,\dim}},$$

where the denominator is the total count of all symbol pairs. The third step is to calculate the observed background frequencies $p_{i,\dim}$ for each symbol using formula

$p_{i,\dim} = \sum_{j=1}^{\dim} \frac{q_{ij,\dim}}{2} + \frac{q_{ii,\dim}}{2}$, and to calculate the expected frequencies of symbol pair i and j occurring in an aligned position $e_{ij,\dim}$ as $2p_{i,\dim}p_{j,\dim}$ for $i \neq j$ and $p_{i,\dim}p_{j,\dim}$ for $i = j$.

We are then able to calculate the substitution score between symbols i and j using formula

Equation 5.2 $$S_{ij,\dim} = \log_2(q_{ij,\dim} / e_{ij,\dim}).$$

This formula gives substitution scores in bit units, consistent with the BLOSUM62 substitution matrix used in the NCBI BLAST search.

5.2.2 Effective Count Calculation For Multiple Sequence Alignment

Our method is used to align two sequence alignments, each with a predicted secondary structure. The input sequence alignments can be generated using different methods or programs (e.g. manual, PSI-BLAST, Pfam) and the distances between sequences in each alignment is unequal. In order to balance the sequence information and make sure the information provided by single distant sequences is not overwhelmed by a large number of similar sequences, each sequence in a group of similar sequences should contribute less to the sequence profile than sequences that are very distant to all others. Therefore, we need to apply a weighting scheme to the input sequence alignment in order to correct for the unequal distances between different sequences.

We use a modified position-specific independent count weighting scheme (Sunyaev, Eisenhaber et al. 1999; Pei and Grishin 2001) to calculate the effective count N_{eff} of each amino acid at different positions. The effective number of sequences N_{eff} is calculated as

Equation 5.3

$$N_{eff} = \frac{\ln\left(1 - \frac{N_{real}}{\dim}\right)}{\ln\left(1 - \frac{1}{\dim}\right)},$$

where dim is 20 for protein sequence alignment and thus the denominator is ln0.95, N_{real} is the average number of different amino acid types per position, and N_{eff} corresponds to the number of random sequences in a random alignment that has the average number of different amino acid types per position equals to N_{real}. When derive the residue content N_{real}^{i} for a given amino acid *i* at a given position, the method only considered a subset of similar sequences that contains *i* at the given position.

This method corrects for the correlation between aligned sequences. When the sequences containing *i* at a given position are identical, N_{eff}^{i} = 1; when the sequences containing *i* at a given position are independent of each other, N_{eff}^{i} equals to the number of these sequences. After having the effective counts, it is easy to calculate the effective frequency of symbol *i* as

Equation 5.4

$$f_{i,\dim} = \frac{N_{eff}^{i}}{\sum_{j=1}^{\dim} N_{eff}^{j}}$$

for a given position in the alignment.

5.2.3 Predicted Secondary Structure Information

Secondary structures of the top sequence of each multiple sequence alignment are predicted using PSIPRED version 2.45(Jones 1999; McGuffin, Bryson et al. 2000).

PSIPRED uses as input the position-specific scoring matrix (sequence profile) generated by PSI-BLAST after three iterations and feeds the profile into a standard feed-forward back-propagation neural network with a single hidden layer to predict secondary structure. The current version of the method takes a consensus prediction from four independently trained neural networks and results in an increased accuracy.

The output of PSIPRED contains three sets. The first set is the predicted secondary structure state, either helix (H), extended strand (E) or coil (C), for each amino acid position in the seed sequence of the PSI-BLAST profile. The second set is the confidence value associated with each predicted state. The confidence value ranges from 1 to 9 with 9 indicates the most confident predictions. The third set of output is the probabilities of C, H, E each occurring at a given amino acid position of the seed sequence. The predicted secondary structure state in the first output set is the one with the highest probabilities for this position.

When using the predicted secondary structure information, we certainly need to take into consideration the confidence value or the probabilities besides the predicted state. After trying different approaches using confidence value and probabilities, we find that using probabilities gives better alignment quality. Therefore, we decide to use probabilities in our method and present the predicted secondary structure as a vector $[p^{\mathrm{H}}, p^{\mathrm{E}}, p^{\mathrm{C}}]$ for each residue position, where $p^{\mathrm{H}}, p^{\mathrm{E}}, p^{\mathrm{C}}$ are the probabilities (normalized to sum up to 1) for helix, strand and coil, respectively. The effective frequencies of secondary structure elements at a given position is hence $f_{i,3} = p^i$, where i equals to H, E and C, respectively.

5.2.4 Estimation Of Target Frequencies

Given a multiple alignment, we need to estimate the target frequencies of each symbol happening at every position. Because of factors like small sample size and prior knowledge about the relationships between symbols, the observed effective frequencies are not good estimations of the target frequencies. Studies have shown that a good approach to estimate target frequency is to mix the effective frequency with a pseudocount frequency. We adopt the pseudocount and target frequency estimation method used in PSI-BLAST and

COMPASS. PSI-BLAST and COMPASS use a data-dependent pseudocount method introduced by Tatusov *et al* (Tatusov, Altschul et al. 1994). This method generates pseudocount $g_{i,\dim}$ for symbol *i* using the observed effective count $f_{i,\dim}$ and the prior knowledge of relationships between symbols that is contained in the substitution matrices $S_{ij,\dim}$ calculated above. This pseudocount is calculated as

Equation 5.5

$$g_{i,\dim} = \sum_{j=1}^{\dim} \frac{f_{j,\dim}}{p_{j,\dim}} q_{ij,\dim} ,$$

where $q_{ij,\dim}$ is the substitution probability between symbol pair *i* and *j* in the corresponding substitution matrices calculated above.

The target frequency is then calculated as a mixture of the effective frequency $f_{i,\dim}$ and pseudocount frequency $g_{i,\dim}$,

Equation 5.6

$$Q_{i,\dim} = \frac{\alpha f_{i,\dim} + \beta g_{i,\dim}}{\alpha + \beta} = \frac{\alpha f_{i,\dim} + \beta \sum_{j=1}^{\dim} f_{j,\dim} \frac{q_{ij,\dim}}{p_{j,\dim}}}{\alpha + \beta},$$

The weight-parameters α and β in the formula are determined empirically as N_c-1 (N_c is the average number of symbol types per position for the input alignment) and 10, respectively, the same as in PSI-BLAST and COMPASS.

5.2.5 Scoring System for Amino Acid And Secondary Structure Profiles

Profile calculation

Profiles are position-specific scoring matrices. They represent the preferences of characteristic amino acids, and in our case secondary structures, of each particular protein family. To calculate the profile scores for each position (column) in an alignment, we use the proved log-odds form (Altschul, Madden et al. 1997; Schaffer, Wolf et al. 1999; Sadreyev and Grishin 2003), $\log \frac{Q_{i,\dim}}{p_{i,\dim}}$, where $Q_{i,\dim}$ is the target frequency of symbol *i*, and $p_{i,\dim}$ is the corresponding background frequency of *i*.

Basic similarity scoring function

In order to align two profiles, we need to have a scoring function to evaluate the similarity between two aligned columns, 1 and 2, each from one of the alignment respectively. We use the basic scoring formula that is extended from COMPASS, which is in turn modified from PSI-BLAST scoring function,

Equation 5.7 $$S_{\dim} = c_1 \sum_{i=1,\dim} n_{i,\dim}^{(1)} \ln \frac{Q_{i,\dim}^{(2)}}{p_{i,\dim}} + c_2 \sum_{i=1,\dim} n_{i,\dim}^{(2)} \ln \frac{Q_{i,\dim}^{(1)}}{p_{i,\dim}},$$

where $Q_{i,\dim}^{(1)}$ and $Q_{i,\dim}^{(2)}$ are the target frequencies of symbol *i* in column 1 and column 2, respectively, $n_{i,\dim}^{(1)}$ and $n_{i,\dim}^{(2)}$ are the effective counts of symbol *i* in column 1 and 2, respectively. c_1 and c_2 are the weighting parameters. They are calculated as

Equation 5.8 $$c_1 = \frac{\sum_i n_{i,\dim}^{(2)} - 1}{\sum_i n_{i,\dim}^{(1)} + \sum_i n_{i,\dim}^{(2)} - 2}, \quad c_2 = \frac{\sum_i n_{i,\dim}^{(1)} - 1}{\sum_i n_{i,\dim}^{(1)} + \sum_i n_{i,\dim}^{(2)} - 2}.$$

This symmetric scoring formula is derived from the probabilities of occurrence of column 1 and column 2 given column 2 and column 1, respectively. c_1 and c_2 are determined so that in the special case when there is only one sequence in a column, the scoring function is transformed to that of PSI-BLAST (for more details, see Ref: COMPASS).

Scoring systems to incorporate secondary structure with amino acid profiles

Two scoring systems are used to incorporate the secondary structure with amino acid profiles. One is a linear combination of the amino acid score (S_{20x20}) and secondary structure score (S_{3x3}) as shown in Equation 5.9. The weighting parameters of the two scoring items sum up to 1 and we only need to optimize one parameter, *w*, which is the weight for secondary structure score. After screening from 0 to 1 with a 0.2 increment, *w* is optimized to 0.2 according to alignment quality. Thus, the optimal weight ratio between amino acid score and secondary structure score is about 4:1.

Equation 5.9

$$S_{20x20+3x3} = (1-w) * S_{20\times 20} + w * S_{3\times 3}$$

The other scoring system we used to incorporate secondary structure and amino acid profiles is to score the combined symbols directly with a 60x60 alphabet. Based on Equation 5.4 and the effective frequency of predicted secondary structure elements (section 5.2.3), the effective frequency of combined symbol *k* is calculated as

Equation 5.10

$$f_{k,60} = \frac{N^{k}_{eff,60}}{\sum_{m=1}^{60} N^{m}_{eff,60}} = \frac{N^{i}_{eff,20} \times f_{j,3}}{\sum_{m=1}^{60} N^{m}_{eff,60}} = \frac{N^{i}_{eff,20} \times f_{j,3}}{\sum_{m=1}^{20} N^{m}_{eff,20}} = f_{i,20} \times f_{j,3}$$

for a given position in the alignment, where symbol *k* is the combination of *i* (a type of amino acid) and *j* (a type of secondary structure element). After having the effective frequencies of combined symbols, we are able to calculate their effective counts $n_{i,60}$ and target frequencies $Q_{i,60}$, and use Equation 5.11 to calculate the combined score S_{60x60}.

Equation 5.11

$$S_{60x60} = c_1 \sum_{i=1,60} n^{(1)}_{i,60} \ln \frac{Q^{(2)}_{i,60}}{p_{i,60}} + c_2 \sum_{i=1,60} n^{(2)}_{i,60} \ln \frac{Q^{(1)}_{i,60}}{p_{i,60}}$$

where $p_{i,60}$ is the background frequency of combined symbol *i*, and c_1 and c_2 are calculated the same way as in Equation 5.8.

Other scoring systems have also been tried out, including linear combination of S_{60x60} and S_{3x3}, and using S_{3x3} alone. Using S_{3x3} alone gives good alignment coverage comparable to $S_{20x20+3x3}$ but bad alignment accuracy. Linear combination of S_{60x60} and S_{3x3} requires more computation tasks but gives similar alignment quality result compared to $S_{20x20+3x3}$. Therefore, both approaches are disregarded.

5.2.6 Estimation Of Statistical Significance Of Optimal Alignments

The local sequence alignment is generated using Smith-Waterman dynamic programming algorithm (Smith and Waterman 1981). The optimal alignment is the one with maximum score. After obtaining the optimal alignment and its associated score, we use a hypothesis testing method to evaluate the statistical significance of the optimal alignment score. The null hypothesis is that the similarity between the two aligned sequences with score S is the result of random chance. The alternative hypothesis is that the similarity between the two aligned sequences with score S is the result of nonrandom reasons and thus is a biologically meaningful similarity.

Determination of statistical parameters

The statistical test requires knowing the distributions of optimal alignment scores for random (non-homologous) alignments. To get the distribution, we generated 10,000 pairs of pseudo sequence alignments with secondary structures composed of randomly selected columns from real alignments. These randomly sampled columns are selected from Pfam 10.0 alignments with effective gap content less than 50%. The 10,000 pairs of pseudo-alignments are then optimally aligned and the scores are calculated. Like PSI-BLAST and COMPASS, the distribution of the optimal scores is well fitted by an extreme value distribution (EVD) with parameters λ and K (Figure 5.2).

According to Altschul and Gish's study (1996, Methods Enzymology), the statistical parameters of EVD depend on search space size, i.e. alignment length (*len*). In our study, the statistical parameters are also found to be dependent on another important property of the alignment, the effective number of sequences in the alignment (N_{eff}). To study these dependencies, we constructed 16 sets of pseudo-alignments of the combinations of 4 alignment lengths (*len* = 100, 200, 300, 500) and 4 effective numbers of sequences (number of sequence = 50, 200, 400, 600, corresponding $N_{eff} \approx$ 10, 15, 17, 18). Each set contains 10,000 pairs of pseudo-alignments constructed as described above. For each given *len* and N_{eff}, the distribution of the random optimal scores is fitted to an EVD with parameters $\lambda(N_{eff}, len)$ and $K(N_{eff}, len)$.

The dependencies of λ and K on N_{eff} and *len* are approximated by planes (Figure 5.3) in the form of $\lambda = a_1 N + b_1 \frac{1}{l} + c_1$ and $K = a_2 N + b_2 \frac{1}{l} + c_2$, where N is the average effective number of sequences of the two profiles in the alignment (i.e. the average of $N_{eff}1$ and $N_{eff}2$), and l is the average alignment length of *len*1 and *len*2. The corresponding coefficients are derived from Figure 5.3 and the dependency relationships are summarized below.

$$\begin{aligned}\lambda_{20x20+3x3} &= -0.00279 \times N + 3.51 \times (1/l) + 0.337 \\ &= -0.00139 \times (N_{eff}1 + N_{eff}2) + 1.76 \times (\frac{1}{len1} + \frac{1}{len2}) + 0.337\end{aligned}$$

$$K_{20x20+3x3} = -0.00643 \times (N_{eff}1 + N_{eff}2) + 8.92 \times (\frac{1}{len1} + \frac{1}{len2}) + 0.299$$

$$\lambda_{60x60} = -0.000228 \times (N_{eff}1 + N_{eff}2) + 1.79 \times (\frac{1}{len1} + \frac{1}{len2}) + 0.273$$

$$K_{60x60} = -0.000220 \times (N_{eff}1 + N_{eff}2) + 4.62 \times (\frac{1}{len1} + \frac{1}{len2}) + 0.679$$

Calculation of E-value

After knowing the distributions of optimal alignment scores for random alignments, we are able to calculate the E-value corresponding to an alignment with optimal score S using the simple formula proposed by Karlin and Altschul (Karlin and Altschul 1990; Altschul, Madden et al. 1997)

Equation 5.12 $$E = Kmne^{-\lambda S}$$

where K and λ are the EVD parameters determined above. And m and n are the lengths of the two profiles in case of pairwise comparison, or the lengths of query and the database in case of database search.

5.3 RESULTS AND DISCUSSION OF THE SUBSTITUTION MATRICES

We calculate three substitution matrices, 20x20, 3x3 and 60x60, at the 80% identity level. Table 5.1 shows the relative entropies and expected scores (means) of these substitution matrices and BLOSUM62. The relative entropy is calculated as $H = \sum_{i,j} q_{ij} S_{ij} = \sum_{i,j} q_{ij} \log_2 \frac{q_{ij}}{p_i p_j}$, and it measures the average information available per position in an alignment (Altschul 1991). From Table 5.1 we can see that the relative entropy and expected score of our 20x20 amino acid substitution matrix are the same as those of BLOSUM62. The relative entropies of 3x3, 20x20 and 60x60 increase in that order, and the expected scores of them decrease in the same order. This phenomenon may be related to the dimensions of the substitution matrices.

Table 5.1 Relative Entropy and Expected Scores of Substitution Matrices

	3x3	20x20	60x60	BLOSUM62
Relative entropy	0.43	0.70	1.12	0.70
Expected score	-0.48	-0.52	-0.88	-0.52

Evolvement of database reflected in substitution matrices

An important parameter in substitution matrix calculation is the sequence identity level at which the database sequences are clustered. The results of our screening for identity levels show that 80% matches best with the BLOSUM62 level (section 5.2.1, Figure 5.1). Therefore, all our substitution matrices are calculated at the 80% identity level. Figure 5.11a shows that the values in our amino acid substitution matrix at 80% identity level (AA80) are comparable to the values in BLOSUM62, while Figure 5.11b shows a systematic difference in the values of our AA62 (our amino acid substitution matrix at 62% identity level) and BLOSUM62.

Since we use the same protocol BLOSUM62 used to calculate our AA62, the only difference between the two substitution matrices is the different sequence data used in the calculation. BLOSUM62 was calculated using the sequence data collected in Blocks database in 1992. With vastly more sequences available in the database since then, the sequences becomes much more divergent.

There is a noticeable trend in the differences between our AA62 scores and BLOSUM62 scores (Figure 5.11b). Compared to BLOSUM62, the scores for substitutions between similar amino acid pairs (the ones with positive substitution scores) are lower in our AA62, while the scores for substitutions between dissimilar amino acid pairs (the ones with negative substitution scores) are higher. This trend is consistent with the changes in the database sequences: the proteins families are including more divergent sequences aligning with each other. Hence in order to obtain comparable values as in BLOSUM62, we now need to use a more similar sequence clustering identity cutoff, 80%, to simulate the degree of sequence diversity more than ten years ago.

Exchangeabilities of Symbol Pairs

According to Equation 5.2, the substitution score between a pair of symbols is positive if their observed frequency is higher than their expected frequency to occur in an aligned position, and negative if their observed frequency is lower than their expected frequency. If a pair of symbols has a very high tendency to substitute each other, they must possess similar chemical properties that enable their similar structural or functional roles in proteins. In addition, their substitution scores will be a large positive value. Therefore, we are able to identify similar amino acid/secondary structure pairs by ranking the corresponding substitution matrices.

All the self-substitutions (diagonal of the substitution matrix) have positive substitution scores. For secondary structure substitution matrix (Figure 5.12a), strand-strand substitution happens most frequently, helix-helix substitution the second, and coil-coil substitution the last but comparable to helix-helix. Other than these self-substitutions, all other substitutions between different types of secondary structure elements are not favorable (have negative substitution scores). For amino acid substitution matrix (Figure 5.12b), the top four similar pairs with substitution scores higher than 1.0 are Tyrosine (Y) and Phenylalanine (F), Tyrosine (Y) and Tryptophan (W), Valine (V) and Isoleucine (I), Lysine (K) and Arginine (R). These amino acid pairs all share similar structural or chemical properties. There are two amino acids that substitute with almost any other amino acids with negative scores (except Cysteine with Alanine). They are Cysteine (C) and Proline (P). These two amino acids have their unique functional or structural roles in proteins. As a result, they are not easily substituted by other amino acids.

Substitution matrices and secondary structure propensities

By definition, the transition probability of the 60x60 combined symbol substitution matrix, $q_{ij,60}$, reflects the dependency between amino acid substitution and secondary structure substitution. In addition, the background frequency of the 60x60 substitution matrix, $p_{i,60}$, reflects the secondary structure propensities for each amino acid type, for $p_{i,60}$ is the frequency of an amino acid type and a secondary structure type occur together. Therefore, we can obtain secondary structure propensities from our calculation of the 60x60 substitution matrix. The distributions of the alpha-helix and beta-strand propensities by amino acid types are shown in Figure 5.13 a & b. Comparisons of the propensities we calculated with the Chou-Fasman propensities (Chou and Fasman 1978) show good correlations with $R^2 = 0.78$ for helix propensities and $R^2 = 0.82$ for strand propensities (Figure 5.13 c&d).

5.4 RESULTS AND DISCUSSION OF PFAM-BASED PERFORMANCE EVALUATION

We evaluate the performance of our methods from two aspects, alignment quality and homology detection ability. As the largest source of accurate semi-automatic multiple sequence alignments and the largest source of automatic structure-based alignments, Pfam sequence alignments and FSSP (Holm and Sander 1996) structure alignments are used in the evaluations. The sequence alignments of Pfam families are used to construct amino acid profiles and the secondary structures are predicted for the top sequences in each alignment using PSIPRED (Jones 1999). We compare the evaluation results of our methods to other methods, including PSI-BLAST (BLAST) (Altschul, Madden et al. 1997) as the most popular tool, COMPASS (Sadreyev and Grishin 2003) as the sequence profile-based method to improve upon, and HHsearch (Soding 2005) and Prof_ss (Chung and Yona 2004) as two methods that also use incorporated sequence profile and secondary structure information.

5.4.1 Evaluation of Alignment Quality

Selection of evaluation data set

500 Pfam 10.0 family pairs with known three-dimensional (3D) structures and structure-based FSSP alignments of sequence identity 14-16% are randomly selected for alignment quality evaluation. The Pfam family pairs are grouped into different identity bins based on the pairwise sequence identities calculated based on the FSSP alignments. Since we want to test the ability of our methods on remotely divergent sequences, we choose to use the family pairs in the identity bin of 14-16%.

Evaluation criteria

When assessing the alignment qualities, the sequence-based alignments (i.e. alignments to be evaluated) are compared to the structure-based FSSP alignments that are used as gold standard. If a pair of residues are aligned the same way in a sequence-based alignment as in the FSSP alignment, it is considered a correctly aligned residue pair and is called a correct match (Sadreyev and Grishin 2003).

The alignment quality is evaluated by coverage, local accuracy and global accuracy. Coverage measures the fraction of structure-based alignment (usually considered containing the longest alignable segments of the two structures) that is reproduced by the sequence-based alignment. It equals to the length between the farmost correctly aligned residue pairs in the sequence-based alignment divided by the length of the structure-based alignment. Local accuracy is also called $Q_{modeler}$ (Yona and Levitt 2002) and measures the percentage accuracy within the aligned region generated by sequence-based alignment. It equals to the number of correct matches divided by the length of the sequence-based alignment. Global accuracy is also called $Q_{developer}$ (Yona and Levitt 2002) and measures the percentage accuracy over the entire structure-based alignment that is reproduced by the sequence-based alignment. It equals to the number of correct matches divided by the length of the structure-based alignment.

Evaluation results and discussion

We compare the alignment quality of our method to other methods in two groups. One is to compare within its own family, including BLAST, PSI-BLAST and COMPASS. We consider these methods of the same family because they use the same basic scoring system and the same statistical significance estimation system. The differences between the methods are the amount of information they use to generate alignments. BLAST uses single-sequence vs. single-sequence, PSI-BLAST uses sequence profile vs. single-sequence, COMPASS uses sequence profile vs. sequence profile, and our methods used sequence profile + secondary structure vs. sequence-profile + secondary structure information. In the other group, we compare our method to HHsearch and Prof_ss, that also used sequence profile + secondary structure information. Figure 5.4a shows that compared to BLAST, PSI-BLAST and COMPASS, our methods (both the 20x20+3x3 and the 60x60 scoring systems) give significantly and substantially improved coverage. Compared to HHsearch and Prof_ss, our method (using the 20x20+3x3 scoring system) also gives the best coverage. Figure 5.4b shows that our methods give best global accuracy in both groups. Figure 5.4c shows that our method (the 20x20+3x3 scoring system) gives slightly worse local accuracy compared to COMPASS.

Two scenarios could cause this worse local accuracy. One is that since our method gives much longer coverage (~1.5 times) compared to COMPASS, alignments generated by our method extend to less similar regions of the two sequences that are more difficult to align. The other scenario is that our method generates random alignments, and the long coverage leads to large global accuracy. To find out which scenario is closer to the truth, we randomly look at the alignments of ~10% of the family pairs. Most of them look like the alignments shown in Figure 5.5 in that the alignments generated by COMPASS are short and the alignments generated by our method contain the COMPASS alignments and extend a lot from the short core regions formed by COMPASS alignments. To further verify our observations, we force the alignments generated by our method and COMPASS to have the same coverage and then compare their accuracy. We use a global version of the dynamic programming algorithm with end gap penalties to generate global alignments for COMPASS and our method, thus the coverage of the alignments are all forced to be 1. Now local accuracy equals global accuracy and there is just one accuracy measure. Figure 5.6 shows that when the coverage are forced to be the same (1.0), our method using the 20x20+3x3 scoring system gives better accuracy than COMPASS. Even using the 60x60 scoring system, our method gives comparable accuracy to that of COMPASS. Therefore, the most plausible reason for our method to have a slightly worse local accuracy is because our alignments extend to less similar regions that are more difficult to align.

5.4.2 Evaluation of Homology Detection Ability

Selection of evaluation dataset and Protocol of sequence similarity searches

To evaluate the homology detection ability, we collect all Pfam 10.0 families that contain at least one sequence belonging to a FSSP family. 1986 Pfam 10.0 families with known 3D structures and available FSSP alignments are selected. An all-against-all search with each of the 1986 families as a query to search against all 1986 families is performed for each of the sequence-based programs, including our methods, COMPASS, HHsearch and Prof_ss. Since these methods all uses multiple sequence profiles, for each query to search against all families, 1986 numbers of searches are performed. But for PSI-BLAST, it uses profile vs. single sequence search. The sequence profile generated from the sequence alignment of each family is used as a query, and the single sequences extracted from the multiple sequence alignment of each other family are transformed to a searchable database. For each query to search against one Pfam family *A*, N numbers of PSI-BLAST searches are performed with N equals to the number of sequences in alignment *A*. The sequence alignments of query against all sequences in family *A* are sorted by E-value. The one with the most significant E-value is used as the hit alignment and E-value for family *A*.

Evaluation criteria

We use similar evaluation criteria as those described in COMPASS method (Sadreyev and Grishin 2003). Our criteria for true positives are based on the consistency between sequence-based alignment and structure-based alignment. The idea is that, if the sequence-based alignment is consistent with the structure-based alignment, it is most likely the result of homology relationship between the query and hit. Thus, we consider the sequence-search hit a true positive if it is consistent with the structure similarity relationship reflected in the FSSP alignment system. If the hit belongs to the same FSSP family as the query, which ensures that they have a structure similarity (DALI) Z-score of greater than 2.0, and if the number of correct matches between the sequence-based alignment and the FSSP alignment of the query and hit is greater than or equals to 5, the hit is considered a true positive. Otherwise, the hit is considered a false positive. Using the number of correct matches of 2 and 15 give similar results.

Evaluation results using ROC curve analysis

Receiver Operating Characteristic (ROC) curve is a popular sensitivity and specificity evaluation technique (McNeil and Hanley 1984). ROC curve analysis is performed to compare the homology search abilities of different methods. The hits or each method are sorted by their E-values in an ascending order. A ROC curve is generated by plotting the numbers of true positives corresponding to each increment in the number of false positives for each method. Figure 5.7a shows the overall ROC curve analysis results. Compared to other methods (HHsearch, COMPASS, Prof_ss, PSI-BLAST), our 20x20+w3x3 approach performs best, but the s60x60 approach performs worse than PSI-BLAST (data not shown). Figure 5.7b shows the ROC curve analysis results of 200 false positives. In this region, our method still performs the best, but its curve is pretty close to that of the COMPASS. Therefore, we need an evaluation method to test if the difference in ROC between these two methods (and between all other methods) is significant or not.

Evaluation results using family-based paired t-test

In order to test if the difference in ROC scores between any two methods is significant or not, we use family-based paired t-test. To perform the test, we calculate the ROC scores for each Pfam family using different sequence search method. We then pair the ROC score of one Pfam family under method one with the ROC score of this family under method two. 1986 pairs of ROC scores are formed for methods one and two and the paired t-test is performed to test if the different in the ROC score means is significant or not at the 5% significant level. Figure 5.7c shows the result of the family-based paired t-test performed within the 200-false positive region (corresponding to that of Figure 5.7b). In this region, at the 5% significant level, the difference between our method and COMPASS is not significant, neither is the difference between COMPASS and HHsearch, but the differences between all other pairs of the 5 methods (our method, COMPASS, HHsearch, Prof_ss and PSI-BLAST) are significant.

5.5 RESULT AND DISCUSSION OF THE 60X60 SCORING SYSTEM APPROACH

We have tried different scoring systems to incorporate secondary structure information with amino acid sequence profiles (section 5.2.5) and decide to use two of the systems, the linear combination 20x20+3x3 approach and the combined symbol 60x60 approach, based on sequence alignment quality comparison. From the evaluation result of homology detection ability (section 5.4.2) we can see that the 20x20+3x3 approach performs better than COMPASS and other methods in the 200-false positive region. However from Figure 5.14 we can see that our method using the 60x60 approach performs worse than COMPASS, the method it is supposed to improve upon. In order to find out the reason why the 60x60 approach fails, we carry out a thorough analysis of this scoring system.

There are three major unique factors in the 60x60 scoring system based on Equation 5.11 and Equation 5.6, (a) the 60x60 substitution matrix, (b) the scoring function and (c) the number of effective count $n_{i,60}$. We study the effects of each factor sequentially. In order to obtain the evaluation result quickly, we randomly choose eight Pfam families as queries to each against the entire Pfam dataset. The ROC curve analysis and family-based paired t-test result of these eight families are shown in Figure 5.15. The evaluation result shows that the 20x20+3x3 approach performs better than COMPASS and the 60x60 approach performs worse than COMPASS (Figure 5.15a) and the differences are all significant (Figure 5.15b), which is consistent with the evaluation result on the entire data set. Thus, we consider it valid to use the randomly chosen eight families to study the effect of the factors.

5.5.1 Study Of The Substitution Matrix Effects

We first consider the substitution matrix used in this scoring system, the 60x60 substitution matrix, which is used to generate data-dependent pseudocount (Equation 5.5) in the scoring system. We start with checking the composition of the dataset we used to derive the 60x60 substitution matrix (Blocks+ 13.0). The distributions of amino acids and predicted secondary elements of the dataset are shown in Figure 5.16. The numbers seems reasonable. A comparison between the amino acid frequencies in the dataset and the Robinson frequencies (Robinson and Robinson 1991) shows strong correlation with $R^2 = 0.90$ (Figure 5.16c). Therefore, we conclude that there is no compositional bias in the dataset we used to derive the substitution matrix.

We then look at the 60x60 substitution matrix itself and find out that its values are very similar to those in an independent matrix. We define an independent 60x60 substitution matrix as one that there is no secondary structure propensity and no dependency between amino acid substitution and secondary structure substitution, i.e. $p_{ax}^{60} = p_a^{20} \times p_x^3$ and $q_{ax\to by}^{60} = q_{a\to b}^{20} \times q_{x\to y}^{3}$, where *a* and *b* represent amino acid type, *x* and *y* represent secondary structure type. Therefore, the values in an independent 60x60 substitution matrix can be deduce to the sums of values in the 20x20 and 3x3 substitution matrices according to Equation 5.13.

Equation 5.13

$$s_{ax\to by}^{60} \equiv \log\frac{q_{ax\to by}^{60}}{p_{ax}^{60}p_{by}^{60}} = \log\frac{q_{a\to b}^{20}}{p_a^{20}p_b^{20}} + \log\frac{q_{x\to y}^{3}}{p_x^3 p_y^3} \equiv s_{ab}^{20} + s_{xy}^{3}$$

Comparison of the values in the original 60x60 substitution matrix and the independent matrix (Figure 5.17) shows a significant correlation between the two (R^2=0.96). Therefore, the 60x60 substitution matrix is indeed very similar to the independent matrix.

To further verify that this similarity has effects on the homology detection ability. We construct the independent substitution matrix based on Equation 5.13 and use this matrix in our scoring system instead of the 60x60 matrix. The ROC curve analysis of the searching result shows that the curve corresponding to the independent matrix almost coincides with that of the 60x60 matrix (Figure 5.18a). And the t-test shows no statistically significant difference between the two (Figure 5.18b). We also carry out a novel hit-rank comparison between the two variations. A hit-rank comparison is defined as such. We sort the hits of each method according to their E-values and record the rank of each hit. The ranks of the same hit by two different methods can then be compared. This way, we can easily identify if two methods give similar search results or not. The advantage of this kind of comparison test is obvious. Because there is not need to determine true or false positives for each hit, it eliminates spurious results. The hit-rank comparison between the independent matrix and the 60x60 substitution matrix shows an exact correlation (R^2=1.0) between the two (Figure 5.18c). Therefore, the 60x60 substitution matrix and the independent matrix not only have similar values, but also have the same effects in homology detection ability.

5.5.2 Study Of The Scoring Function Effects

In order to find out the difference between the two scoring systems 20x20+3x3 (Equation 5.9) and 60x60 (Equation 5.11), we decide to decompose the scoring functions of the two. To simplify the decomposition process and clarify the results, we take only the scoring items contributed by one amino acid, Alanine (A), and one secondary structure, Helix (H), in part one of the symmetric scoring function (Equation 5.7), $c_1 \sum_{i=1,\dim} n_{i,\dim}^{(1)} \ln \frac{Q_{i,\dim}^{(2)}}{p_{i,\dim}}$. Thus, the 20x20+3x3 scoring function is decomposed to

Equation 5.14 $$c_1 \ln \left[\left(\frac{Q_A}{p_A} \right)^{n_A(1-w)} \left(\left(\frac{Q_h}{p_h} \right)^{n_A p_h^{PSI}} \left(\frac{Q_e}{p_e} \right)^{n_A p_e^{PSI}} \left(\frac{Q_c}{p_c} \right)^{n_A p_c^{PSI}} \right)^{w} \right],$$

and assuming $\frac{Q_{Ah,60}}{p_{Ah,60}} = \frac{Q_{A,20}}{p_{A,20}} \times \frac{Q_{h,3}}{p_{h,3}}$, the 60x60 scoring function is decomposed to

Equation 5.15 $$c_1 \ln\left[\left(\frac{Q_A}{p_A}\right)^{n_A}\left(\frac{Q_h}{p_h}\right)^{n_A p_h^{PSI}}\left(\frac{Q_e}{p_e}\right)^{n_A p_e^{PSI}}\left(\frac{Q_c}{p_c}\right)^{n_A p_c^{PSI}}\right],$$

where n_A is the effective count of alanine, p_h^{PSI}, p_e^{PSI} and p_c^{PSI} are the PSI-BLAST predicted probabilities of helix, strand and coil, respectively. Comparison of Equation 5.14 and Equation 5.15 shows that the only difference between the two is the parameter w, which is the weight of the S_{3x3} item in the 20x20+3x3 scoring system ($S_{20x20+3x3} = (1-w)*S_{20\times 20} + w*S_{3\times 3}$). Therefore, the 60x60 scoring function is equivalent to the 20x20+3x3 scoring function with equal weight (both weights of S_{20x20} and S_{3x3} equal to 1).

To verify the deduced equivalence between the 60x60 and $S_{20\times 20} + S_{3\times 3}$ approaches, we carry out a search using the 20x20+3x3 scoring system with both weights equal to 1 and compare the results with that of the 60x60's. As expected, the ROC curve of the two variations overlap with each other (Figure 5.19a), the t-test shows no statistically significant difference between the two (Figure 5.19b), and the hit-rank comparison shows an excellent correlation (R^2=0.996) between the two (Figure 5.19c). These experimental results show that the 60x60 scoring system is indeed equivalent to the 20x20+3x3 scoring system with equal weights, 1.

To think from another direction, we can add weight w and (1-w) to Equation 5.15 so that it becomes the same as Equation 5.14. This way, we can construct a new 60x60 scoring system that is equivalent to the 20x20+3x3 scoring system. As it turns out, adding weight to the 60x60 approach requires adding S_{3x3} in as well. By reversing the deduction process above, we deduce that the scoring system

Equation 5.16 $$(1-w)*S_{60x60} - (1-2w)*S_{3x3}$$

should be equivalent to the 20x20+3x3 system, $S_{20x20+3x3} = (1-w)*S_{20\times 20} + w*S_{3\times 3}$. To verify our thought, we use Equation 5.16 to do a search and compare the results with that of the 20x20+3x3 system. Indeed, the ROC curve shows overlapping of the two methods

(Figure 5.20a), the t-test shows no statistically significant difference (Figure 5.20b), and the hit-rank comparison shows a strong correlation ($R^2=0.91$) between the two (Figure 5.20c).

In short, the analysis of the scoring functions reveals that the fundamental difference between the 60x60 and the 20x20+3x3 scoring systems is the difference in their weights assigned to the amino acid score ($S_{20\times20}$) and the secondary structure score ($S_{3\times3}$). The 20x20+3x3 scoring system has a weight ration of 4:1 for $S_{20\times20}$: $S_{3\times3}$, while in the 60x60 scoring system this ratio is 1:1.

5.5.3 Study Of The Effective Count Effects

In order to study the effects of the number of effective count $n_{i,60}$, we first look at the distribution of $n_{i,60}$ (Figure 5.21a). It ranges from 0.6 to 21. If the number of effective count has effects on the homology detection ability of the 60x60 scoring system, different number of effective count would result in different detection performance. Thus, we can use a stratifying method to test the effects of the number of effective count.

The inter-quartile range of $n_{i,60}$ is 6-14. We first divide the Pfam families into three categories: $n_{i,60} < 6$, $6 \le n_{i,60} \le 14$, and $n_{i,60} > 14$, then use them as queries to execute searches. The search results are evaluated using ROC curve analysis. From Figure 5.21b we can see, the performance of 60x60 is always the worst, and the relative performance of 20x20+3x3, COMPASS and 60x60 do not change between the three categories. Therefore, we conclude that the number of effective count does not seem to affect the performance of the 60x60 scoring system.

In summary, the reason why the 60x60 scoring system approach does not perform as well as the 20x20+3x3 approach and COMPASS is mainly because (a) the 60x60 substitution matrix is very similar to an independent matrix and thus does not contain enough information for homology detection purposes, and (b) the amino acid information $S_{20\times20}$ and the secondary structure information $S_{3\times3}$ have a fixed relative weight (1:1) in the 60x60 scoring system, which is improper (unoptimal) for the homology detection purpose. The number of effective count in the scoring system does not seem to affect the performance.

5.6 APPLICATIONS

With improved alignment quality and structure similarity detection ability, our method is able to provide many applications to the structure modeling field, including detecting template of distant-similarity and providing better alignment for structure modeling. In addition, our method can be used to detect distant similarities between families with known-structures, which is useful for protein structure classification and for understanding of protein sequence-structure-function relationships.

Figure 5.8 shows an example of the improved distance-similarity detection ability and better alignment quality helps to identify and confirm homology relationship between proteins that are otherwise difficult to identify using sequence profile-based method (COMPASS). We already know that both eukaryotic peptide chain release factor subunit 1 C-terminal domain (ERF1) (1dt9:A277-A422) and RNA ribose methyltransferase N-terminal domain (RNArm) (1ipa:A1-A105) are homologous to ribosomal protein L30e through transitive PSI-BLAST, therefore, ERF1 and RNArm should be homologous to each other. However, a database search using sequence profile-based method, COMPASS, fails to find a significant similarity between the two (E-value only 0.7) and the alignment generated only covers one helix and one and half strands (pink regions in Figure 5.8 a&b). The database search using our method is able to find a significant similarity between the two protein domains with an E-value of 7.3e-16 and therefore is able to infer homology relationship between the two. The alignment generated by our method covers the entire length of the two domains and is correctly aligned in 6 out of 7 secondary structures of the two domains.

We apply our method to predict new similarities between Pfam 10.0 families by comparing all Pfam families with known 3D structures in an all-against-all fashion. We select hits that have significant E-values and verify them by structure-based comparisons (number of correct matches). Our method is able to reliably predict similarities between 1809

Pfam family-pairs, which is about 16 times of what COMPASS predicts, and about 28 times of what PSI-BLAST predicts (Table 5.2).

Table 5.2 Number of Similar Pfam family-pairs Predicted

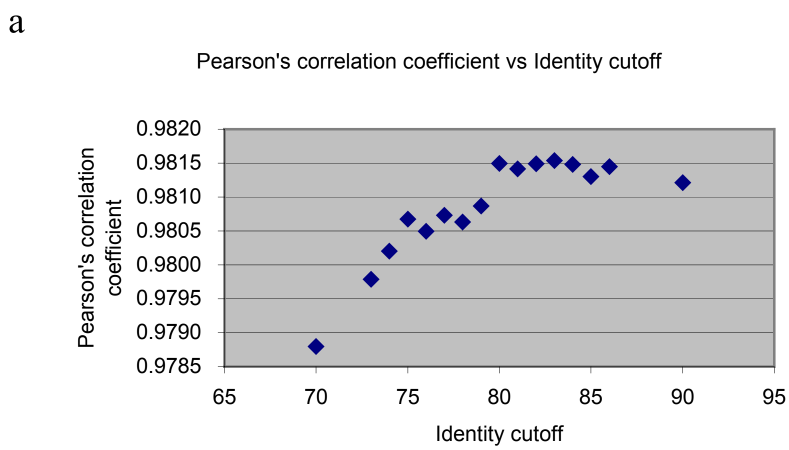
	Number of Pfam family-pairs reliably predicted
20x20+3x3	1809
COMPASS	111
PSI-BLAST	65

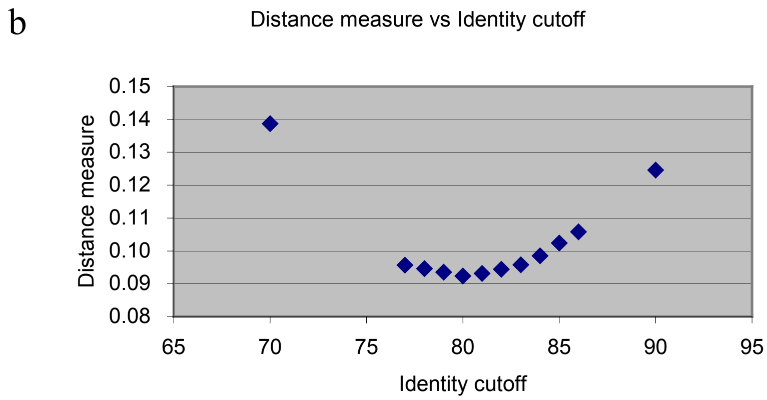
Many interesting examples are found among the new similarities uniquely predicted by our method. Figure 5.9 shows an example of new similarity detected between different SCOP superfamilies. Because no significant sequence similarities detected between the two domains before, Pfam families, heavy-metal-associated domain (HMA) and hydroxymethylglutaryl-CoA reductase (HMG-CoA_red), belongs to two different SCOP (version 1.69) superfamilies, heavy metal-associated domain (d.58.17) and NAD-binding domain of HMG-CoA reductase (d.58.20). Both domains possess the ferredoxin-like fold with a secondary structure arrangement of $(\beta\alpha\beta)_2$. The structural similarities between the two domains were only identified by visual inspection (the same SCOP Fold). Our method is able to identify a significant similarity between the two domains with an E-value of 2.44e-17 and correctly aligns 4 out of the 6 secondary structures in both domains (the red regions in Figure 5.9). Both domains can serve as a perfect structure template for each other. Figure 5.10 shows an example of new similarity detected between different SCOP folds. Pfam families ACT domain and eukaryotic initiation factor 4E (IF4E) belongs to different SCOP folds, ferredoxin-like fold (d.58) and translation initiation factor eIF4e fold (d.86), and were considered to possess different structural folds. This usually means that the two domains cannot be used as structure template for each other. However, our method successfully finds the structurally similar regions in the two domains. ACT domain is a ferredoxin-like fold domain, while the IF4E domain also possesses the ferredoxin-like fold with insertions of other secondary structures at the N- and C-termini (see Figure 5.10). Our method correctly identifies the structurally similar parts of the two domains and correctly aligned $4\frac{1}{2}$ out of 6 secondary structures in the smaller ACT domain. Thus the bigger IF4E domain can serve as a perfect structure template for the ACT domain.

5.7 CONCLUSIONS

A protein sequence alignment and similarity search algorithm has been developed by means of incorporating sequence profile and predicted secondary structure information. We constructed substitution matrices of amino acids and secondary structure elements based on updated sequence database and utilized them in our newly developed scoring system. Statistical significance of the resulting alignments are estimated based the modeled random score distributions. Comparisons to other programs (e.g. PSI-BLAST, COMPASS, Prof_ss) on a PFAM-based performance evaluation system show that our program provides improved template detection ability and generates better quality sequence alignments. Applying our program to PFAM 10.0 families reveals many (16-28 times) previously unrecognized similarities between families. Additionally, we explored different approaches to incorporate predicted secondary structure information with sequence profile and made an effort to understand why some approach does not work from various perspectives.

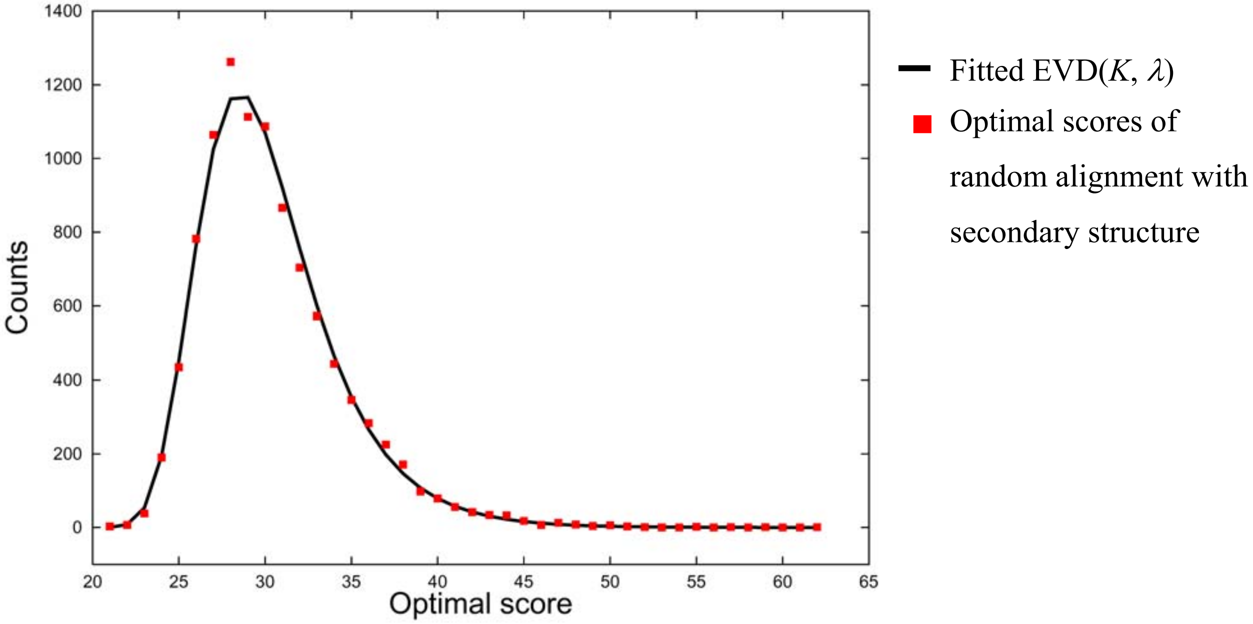
Figure 5.1 My Amino Acid Substitution Matrix AA80 Matches BLOSUM62





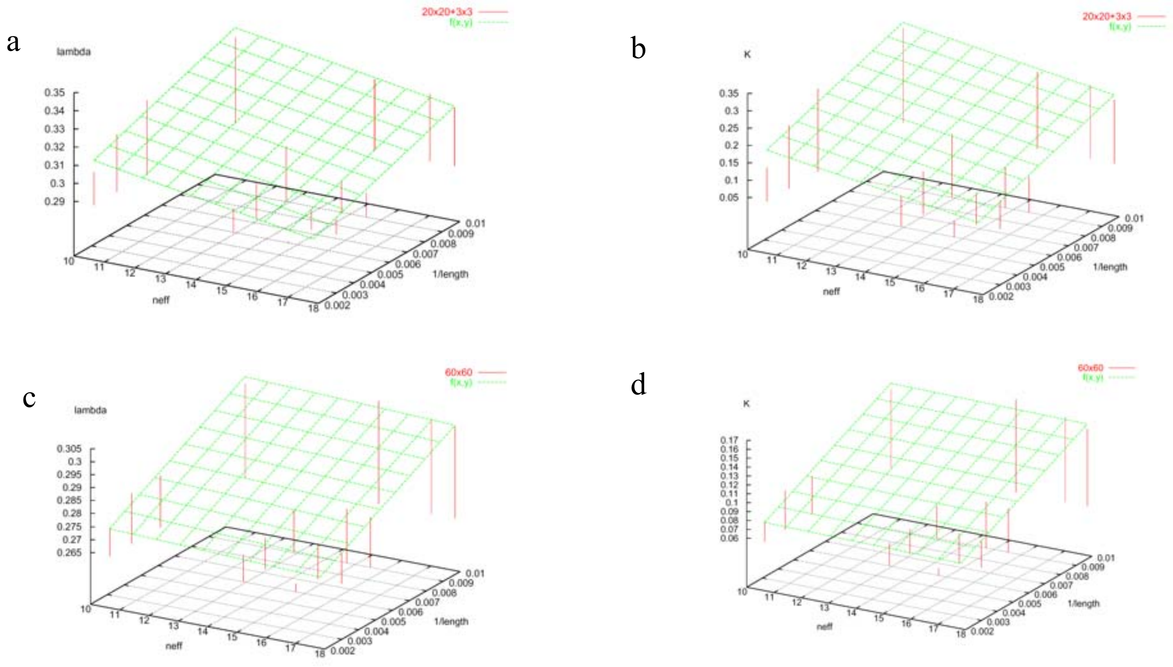
(a) Plot of Pearson's correlation coefficients vs. sequence identity level cutoffs for clustering. Pearson's correlation coefficients are calculated for values in BLOSUM62 and values in our amino acid substitution matrices at different sequence identity levels. This plot shows that the highest correlation coefficient occurs at identity level 80%. (b) Plot of distance measure vs. sequence identity level cutoffs for clustering. The distance measure assesses the dissimilarities between BLOSUM62 and our amino acid substitution matrices at different identity levels. This plot shows that the least dissimilar (=most similar) point occurs at identity level 80%. The distance measure is calculated as the average of d, where d is the vertical offset to line y=x.

Figure 5.2 Distribution of Optimal Random Scores Fitted to EVD



Distribution of optimal scores for 10,000 pairs of pseudo sequence alignments with secondary structures composed of randomly sampled columns from Pfam alignments. The score distribution is generated using the 60x60 scoring system with pseudo-alignments of length 100 and number of sequences 200. The best-fitted Extreme Value Distribution (EVD) is plotted against the data. The chi-square goodness-of-fit test generates a chi-square value of 45.5 with a degree of freedom of 41, which corresponds to a p-value of 0.29.

Figure 5.3 Dependency of Lambda and *K* on Alignment Length and Effective Sequence Number



(a) Dependency of Extreme Value Distribution (EVD) parameter lambda on alignment length (length) and effective sequence number (neff) for 20x20+3x3 scoring system. The tops of the red lines indicate the values of lambdas at the combinations of lengths 100, 200, 300, 500 and neffs 10, 15, 17, 18. The green plane indicates the fitted plane of lambda as a function of length and neff. (b) Dependency of EVD parameter *K* on length and neff for 20x20+3x3 scoring system. The tops of the red lines indicate the values of *K*s at the combinations of lengths 100, 200, 300, 500 and neffs 10, 15, 17, 18. The green plane indicates the fitted plane of *K* as a function of length and neff. (c) Dependency of EVD parameter lambda on length and neff for 60x60 scoring system. The color scheme is the same as in (a). (d) Dependency of EVD parameter *K* on length and neff for 60x60 scoring system. The color scheme is the same as in (b).

Figure 5.4 Evaluation Results of Alignment Quality

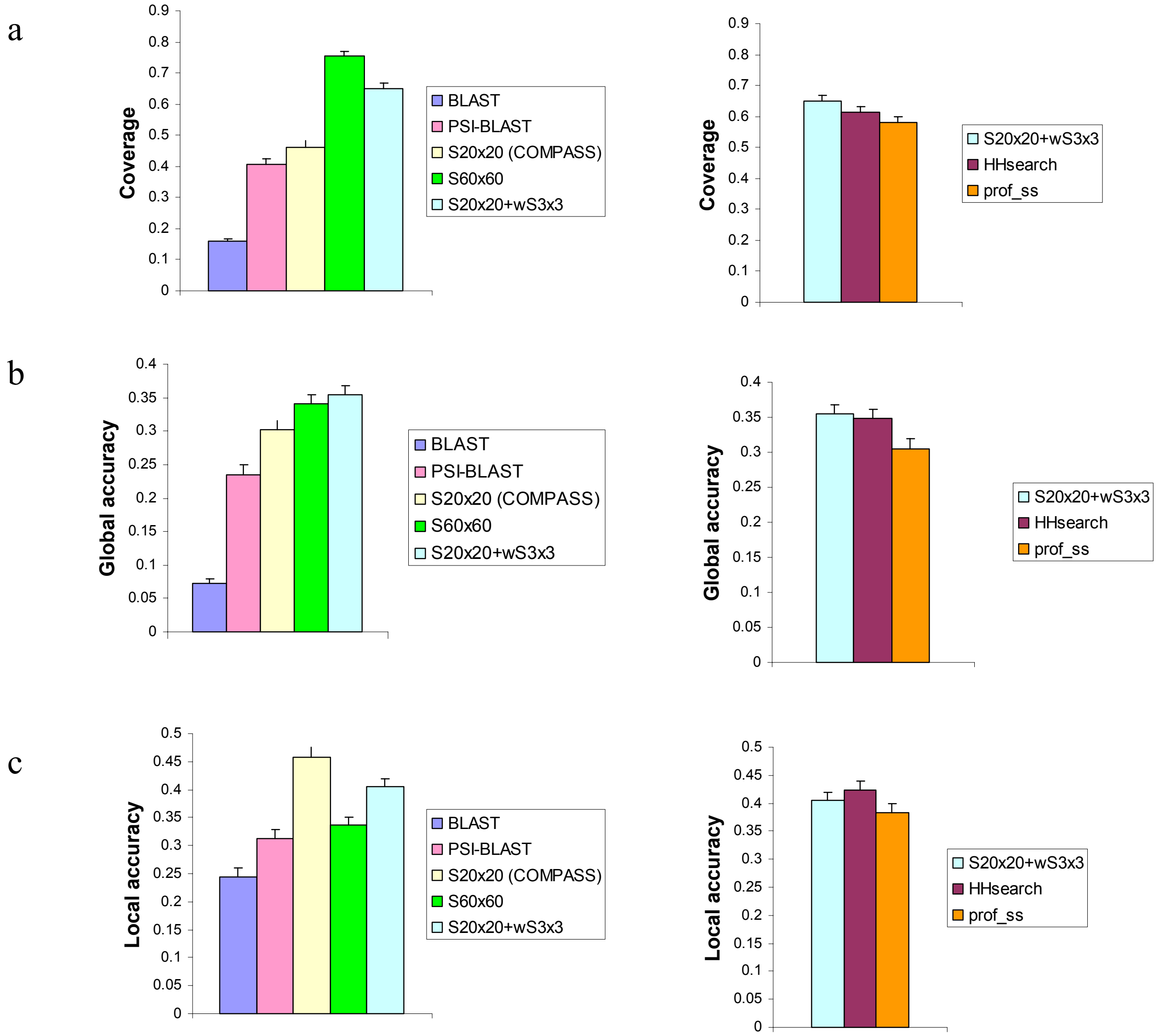
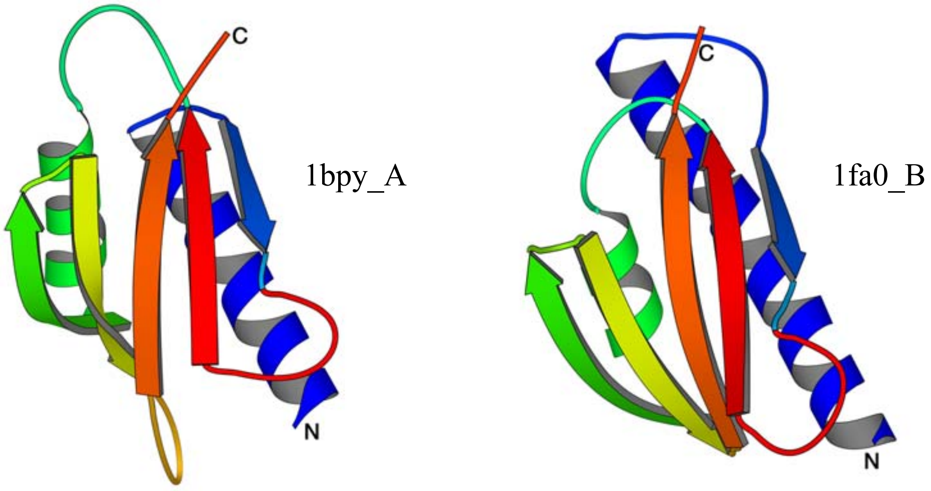


Figure 5.5 Example of Alignments Generated by Our Method and COMPASS

a

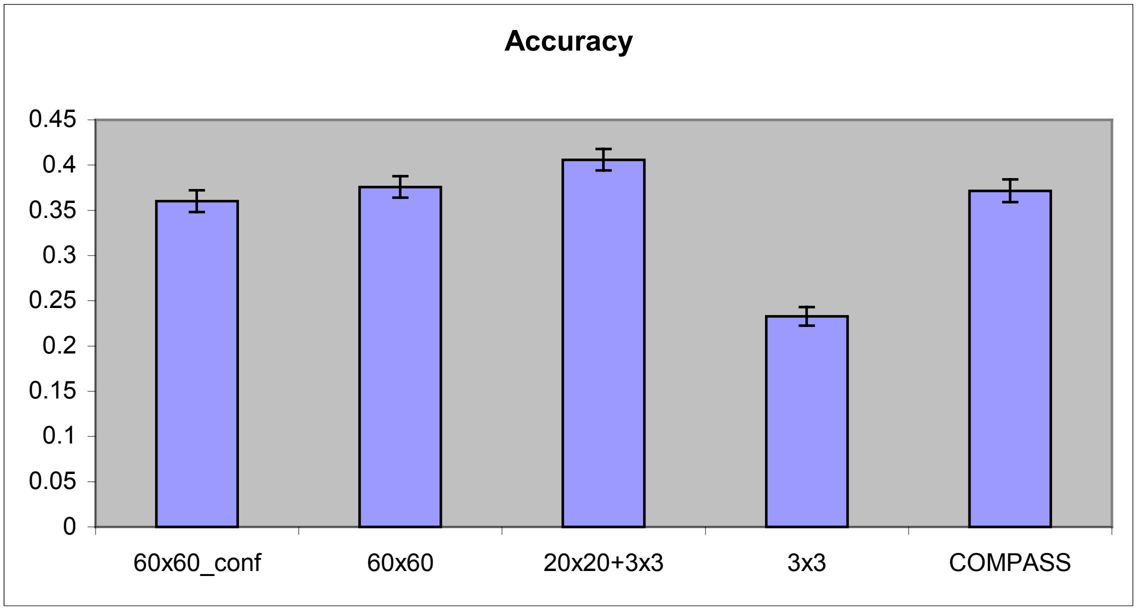
```
Pred:          HHHHHHHHCCCCCHHHHHHH-----HHHHHHHHHHHHHHHCCCCCCCCHHHH-HHHHHHH
1bpyA   100    SAARKFVDEGIKTLEDLRKN=====EDKLNHHQRIGLKYFGDFEKRIPREEM=LQMQDIV
               + +++++       ++++++     ++++++                 + ++ +++++++
1fa0B   23     KLNDSLI=======QELKKEGSFETEQETAN=================RVQVLKILQELA
               HHHHHHH-------HHHHHCCCCCCHHHHHH-----------------HHHHHHHHHHHH

               HHHHHHH------------CCCCEEEEECCCHHHHHHCC------CEEEEEECCCCCCCH
1bpyA          LNEVKKV============DSEYIATVCGSFRRGAESSG======DMDVLLTHPSFTSES
               ++ ++++             ++ ++++ ++++      +      ++++++++ ++  +++
1fa0B          QRFVYEVSKKKNMSDGMARDAGGKIFTYGSYRL=====GVHGPGSDIDTLVVVPKHVTRE
               HHHHHHHHHHHHHHHHCCCCCCEEEEEECCCCC-----CCCCCCCCEEEEEECCCCCCHH

               HHHHHHHHHHHHHHHHCCCCEEE----------ECCCCEEEEEEEECCCCCCCCCCCCEE
1bpyA          TKQPKLLHQVVEQLQKVHFITDT==========LSKGETKFMGVCQLPSKNDEKEYPHRR
                     ++++ +++++++   +           +     +++              + ++
1fa0B          ====DFFTVFDSLLRERKELDEIAPVPDAFVPII=====KIK=============FSGIS
               ----HHHHHHHHHHHHCCCCCCEEEEECCCCCEE-----EEE-------------ECCEE

               EEEEE-------ECCCCCCCEE-----EC---------CCC--CHHHHHH----------
1bpyA          IDIRL=======IPKDQYYCGV=====LY=========FTG==SDIFNKN==========
               +++ +       ++ +    ++     +            +    ++  +
1fa0B          IDLICARLDQPQVPLSL===TLSDKNLLRNLDEKDLRALNGTRVTDEILELVPKPNVFRI
               EEEEECCCCCCCCCCCC---CCCHHHHHHHHHHCCCCCCCCHHHHHHHHHHHCCCCCHHH

               ----HHHHHHHCCCEECC
1bpyA          ====MRAHALEKGFTINE
                   +++ +++++  +++
1fa0B          ALRAIKLWAQRRA==VYA
               HHHHHHHHHHHCC--CCC
```

b



(a) The sequence alignment of 1bpy_A and 1fa0_B generated by our method (20x20+3x3 scoring system). The red regions are the alignment generated by COMPASS. (b) The structure diagrams of 1bpy_A (DNA polymerase beta, catalytic fragment) and 1fa0_B (poly(A) polymerase N-terminal catalytic domain). The red regions in the diagrams (a loop followed by a beta-strand) are the correspondingly only-aligned regions by COMPASS in the sequence alignment.

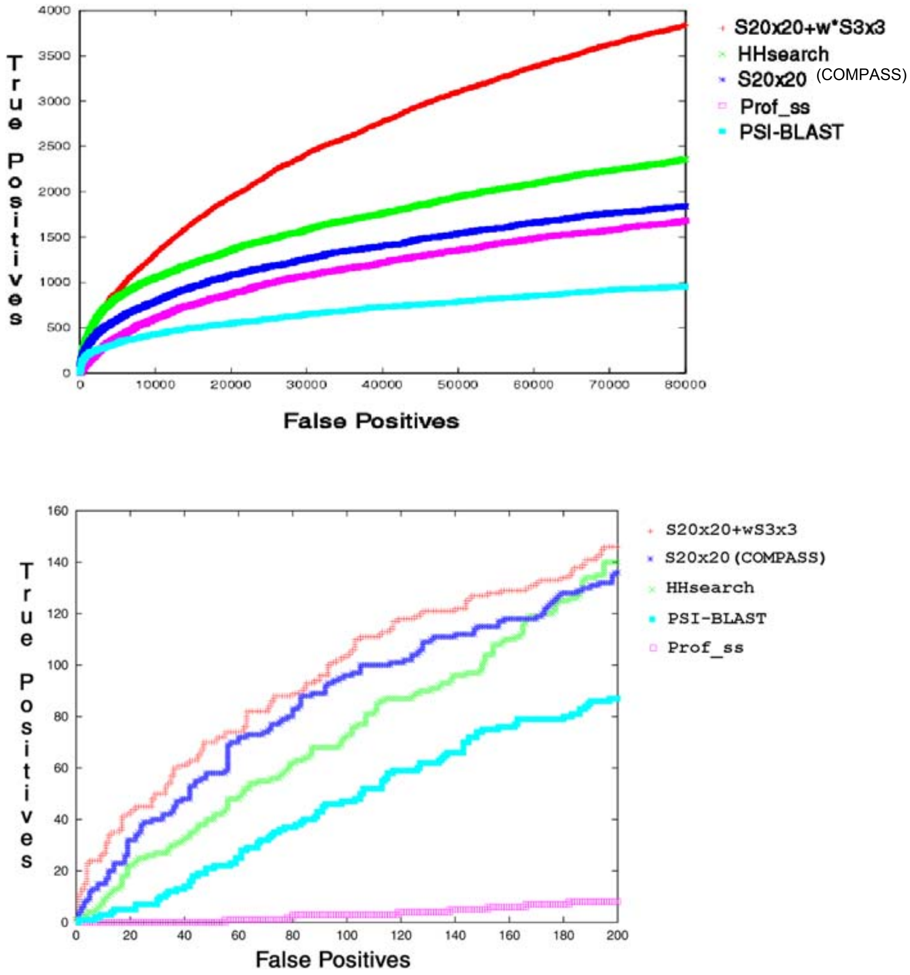
Figure 5.6 Better Accuracy Than COMPASS when Coverage Equals 1



We use global alignment algorithm with end gap penalty to force the coverage of alignments to be 1. As a result, local accuracy equals global accuracy. When having the same coverage (1.0), our method using the 20x20+3x3 scoring system gives better accuracy than that of COMPASS, and the 60x60 scoring system gives comparable accuracy than that of COMPASS. The “60x60_conf” and “3x3” are other scoring systems we tried.

Figure 5.7 Evaluation Results of Homology Detection Ability

a.



b.

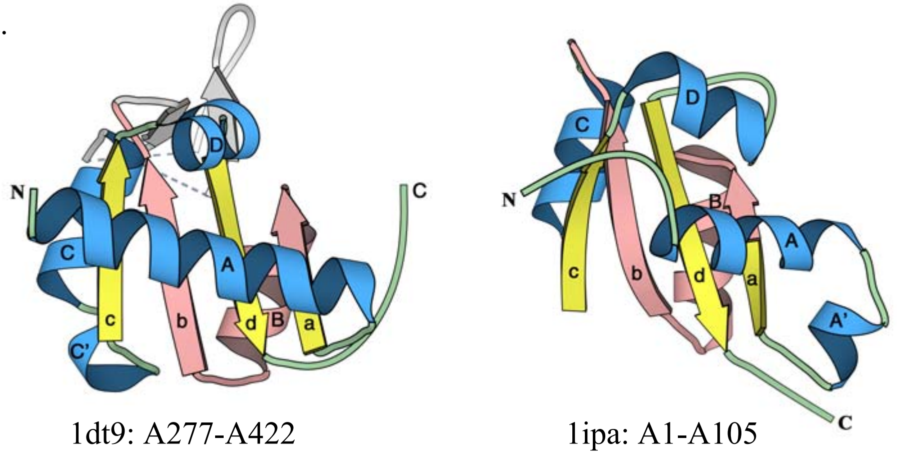
c.

	20x20+3x3	COMPASS	HHsearch	PSI-BLAST	Prof_ss
20x20+3x3	*				
COMPASS	–	*			
HHsearch	+	–	*		
PSI-BLAST	+	+	+	*	
Prof_ss	+	+	+	+	*

(a) Overall ROC analysis results showing the homology detection ability of various programs. Different colors indicate different programs. (b) ROC analysis results of false positive 200 (FP200). (c) Family-based t-test results of ROC FP200. At the 5% significant level, + indicates the difference is significant, – indicates the difference is not significant. * means self-comparison, which is meaningless.

Figure 5.8 Example of Improved Homology Detection Ability and Alignment Quality

a.



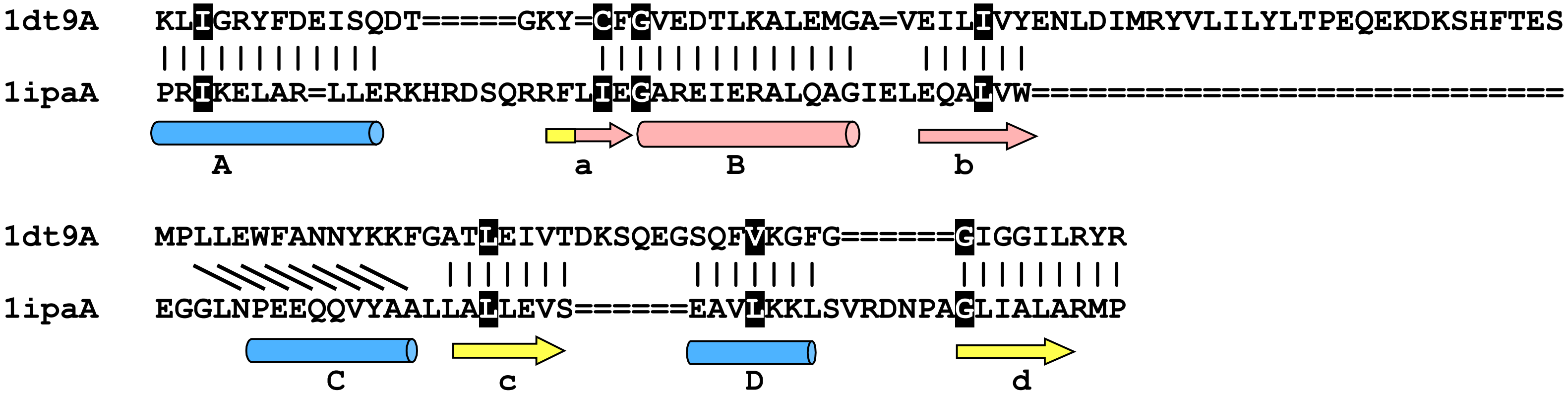
1dt9: A277-A422

1ipa: A1-A105

(a) Structure diagrams of eukaryotic peptide chain release factor subunit 1 C-terminal domain (1dt9:A277-A422) and RNA ribose methyltransferase N-terminal domain (1ipa:A1-A105). (b) Database search result using our method with 1dt9 as query. The pink region corresponds to the alignment generated by database search using COMPASS and the E-value corresponding to COMPASS search is 0.7.

b.

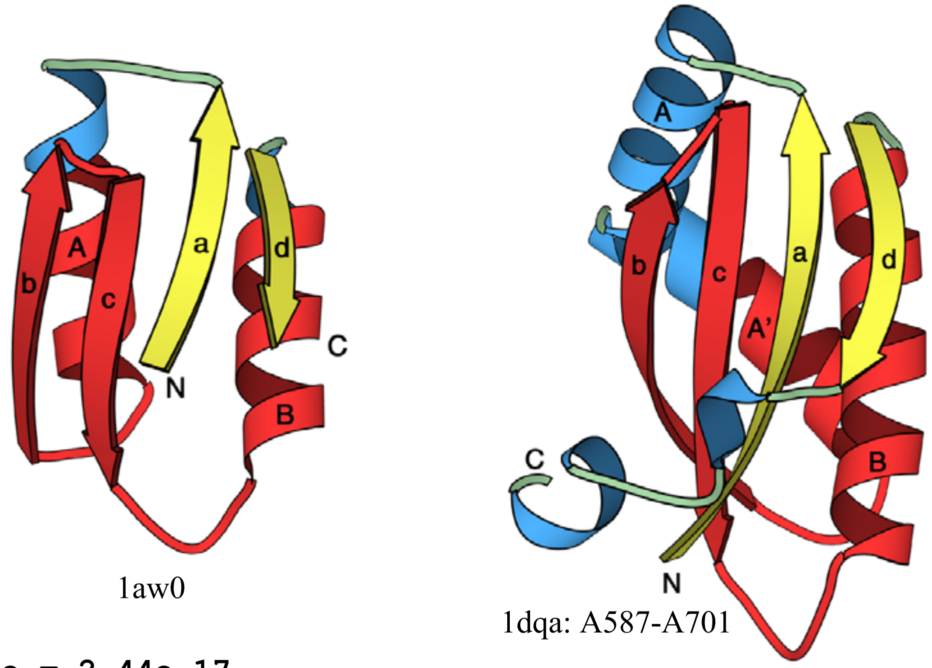
Evalue = 7.3e-16, database size = 3.3x10^5



```
1dt9A    KLIGRYFDEISQDT=====GKY=CFGVEDTLKALEMGA=VEILIVYENLDIMRYVLILYLTPEQEKDKSHFTES
         |||||||||||||          ||||||||||||||| ||||||
1ipaA    PRIKELAR=LLERKHRDSQRRFLIEGAREIERALQAGIELEQALVW============================

1dt9A    MPLLEWFANNYKKFGATLEIVTDKSQEGSQFVKGFG======GIGGILRYR
            \\\\\\\\    |||||||      |||||||       |||||||||
1ipaA    EGGLNPEEQQVYAALLALLEVS======EAVLKKLSVRDNPAGLIALARMP
```

Figure 5.9 Example Of New Similarity Detected Between SCOP Superfamilies

a.



1aw0

1dqa: A587-A701

New similarities detected using our method (20x20+3x3) between two families belonging to different SCOP superfamilies. (a) Structure diagrams of the representative domains of the two families. 1aw0 is the 4th metal binding domain of Menkes copper-transporting ATPase, which belongs to the Pfam family of heavy-metal-associated domain. 1dqa is the NAD-binding domain of HMG-CoA reductase, which belongs to the Pfam family of hydroxymethylglutaryl-CoA reductase. The red regions in the structure diagrams corresponding to the aligned red regions in the alignment (b).

b.

Evalue = 2.44e-17

```
1aw0    QSIEGVISKKP===GVKSIRVSLANSNG==TVEYDPLLTS========PETLRGAIED
                      |||||||||   \\  |||||             ||||||||||
1dqaA   AVIKEAFDSTSRFARLQKLHTSIAGRNLYIRFQSRSGDAMGMNMISKGTEKALSKLHE
```

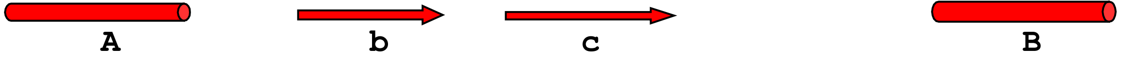
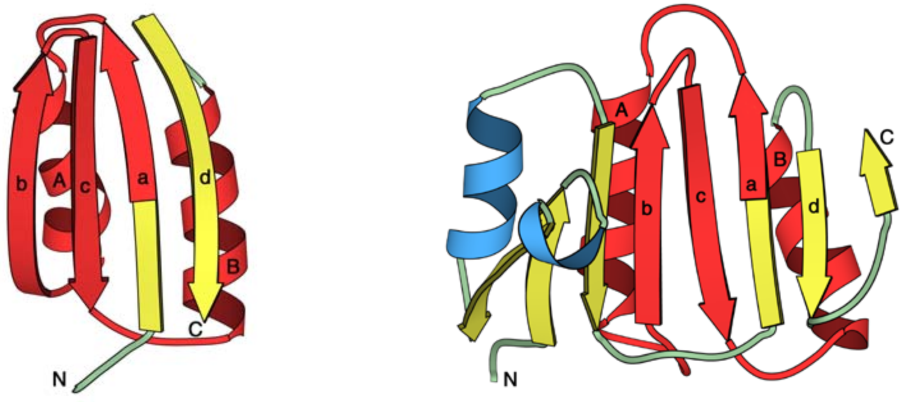


Figure 5.10 Example Of New Similarity Detected Between SCOP Folds

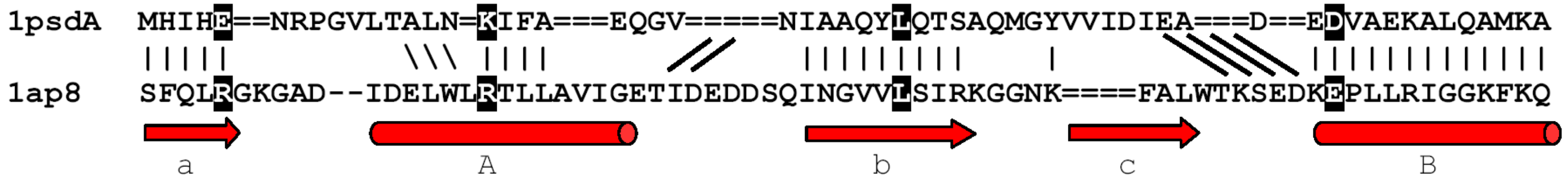
a.



1psd:A327-A410 1ap8

(a) Structure diagrams of C-terminal regulatory domain of phosphoglycerate dehydrogenase (1psd:A327-A410) that belongs to the Pfam family of ACT domain, and an representative structure (1ap8) of the Pfam family of eukaryotic initiation factor 4E. (b) The sequence alignment generated using our method (20x20+3x3 approach). The red regions in the structures are corresponding to the red secondary structure highlighted in the alignment.

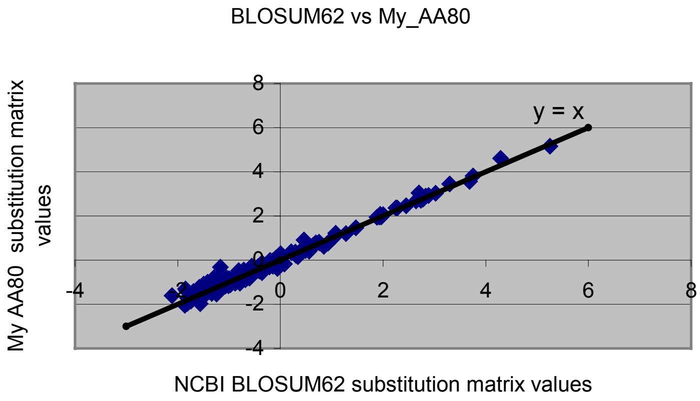
b.



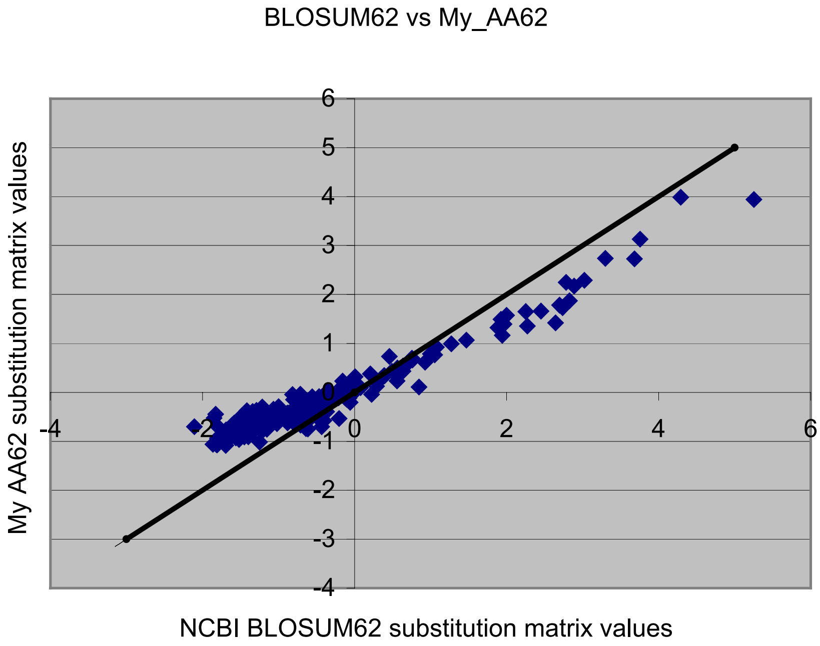
```
E-value = 7.38e-09

1psdA  MHIHE==NRPGVLTALN=KIFA===EQGV=====NIAAQYLQTSAQMGYVVIDIEA===D==EDVAEKALQAMKA
       |||||         \\\ ||||        //   |||||||||    |       \\\\  |||||||||||||
1ap8   SFQLRGKGAD--IDELWLRTLLAVIGETIDEDDSQINGVVLSIRKGGNK====FALWTKSEDKEPLLRIGGKFKQ
```

Figure 5.11 Comparison of BLOSUM62 and Our Amino Acid Substitution Matrix

a.



b.



(a) The values in our amino acid substitution matrix clustered at 80% sequence identity level (AA80) matches the values in BLOSUM62 substitution matrix. (b) The values in our amino acid substitution matrix clustered at 62% sequence identity level (AA62) are systematically different from the values in BLOSUM62.

Figure 5.12 The Three Substitution Matrices We Calculated

a.

	H	E	C
H	0.932		
E	-2.147	1.554	
C	-1.186	-0.489	0.852

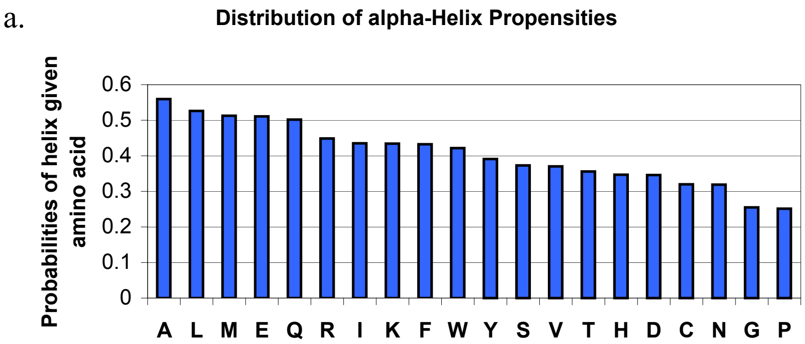
b.

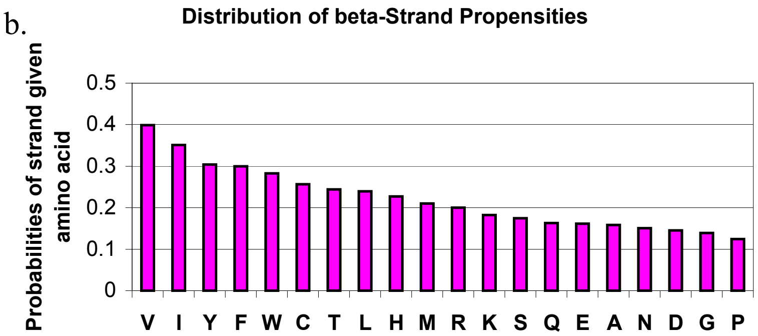
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1.94																			
R	-0.75	2.72																		
N	-0.77	-0.23	2.90																	
D	-0.90	-0.73	0.67	2.94																
C	0.02	-1.44	-0.95	-1.63	4.63															
Q	-0.43	0.44	0.06	-0.04	-1.29	2.69														
E	-0.43	-0.09	-0.12	0.82	-1.72	0.78	2.47													
G	-0.18	-1.22	-0.29	-0.78	-1.17	-0.99	-1.14	2.76												
H	-0.93	-0.02	0.39	-0.27	-1.07	0.33	-0.37	-1.09	3.82											
I	-0.71	-1.37	-1.54	-1.98	-0.77	-1.22	-1.52	-2.04	-1.37	2.05										
L	-0.75	-1.18	-1.46	-1.90	-0.79	-0.94	-1.42	-1.97	-1.12	0.78	1.95									
K	-0.54	1.06	0.10	-0.23	-1.48	0.61	0.40	-0.96	-0.13	-1.31	-1.20	2.37								
M	-0.37	-0.89	-1.08	-1.63	-0.54	-0.29	-1.11	-1.50	-0.82	0.54	0.94	-0.80	3.04							
F	-0.99	-1.43	-1.42	-1.87	-0.76	-1.24	-1.69	-1.70	-0.43	0.03	0.38	-1.53	0.29	3.02						
P	-0.49	-0.84	-0.84	-0.68	-1.46	-0.60	-0.50	-1.16	-0.91	-1.40	-1.38	-0.51	-1.35	-1.47	3.56					
S	0.40	-0.49	0.22	-0.20	-0.30	-0.18	-0.28	-0.27	-0.45	-1.28	-1.29	-0.25	-0.78	-1.16	-0.26	2.07				
T	-0.14	-0.56	-0.09	-0.61	-0.35	-0.28	-0.47	-1.06	-0.62	-0.58	-0.82	-0.30	-0.38	-0.98	-0.56	0.79	2.36			
W	-1.14	-0.98	-1.32	-1.61	-1.08	-1.11	-1.43	-1.52	-0.31	-0.78	-0.47	-1.21	-0.43	0.92	-1.37	-1.15	-1.15	5.16		
Y	-1.04	-0.78	-0.80	-1.37	-0.87	-0.80	-1.16	-1.69	0.65	-0.67	-0.46	-0.94	-0.40	1.46	-1.28	-0.97	-0.89	1.22	3.45	
V	-0.14	-1.22	-1.39	-1.74	-0.25	-1.04	-1.20	-1.78	-1.26	1.21	0.32	-1.13	0.17	-0.28	-1.13	-0.90	-0.18	-0.98	-0.72	1.93

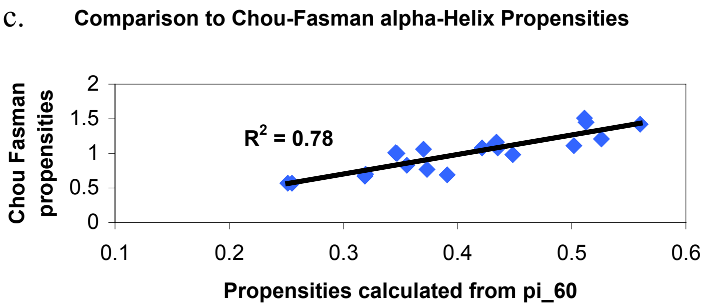
c. http://iole.swmed.edu/~yqi/sub_mtx/sub_mtx.80.AS.xls

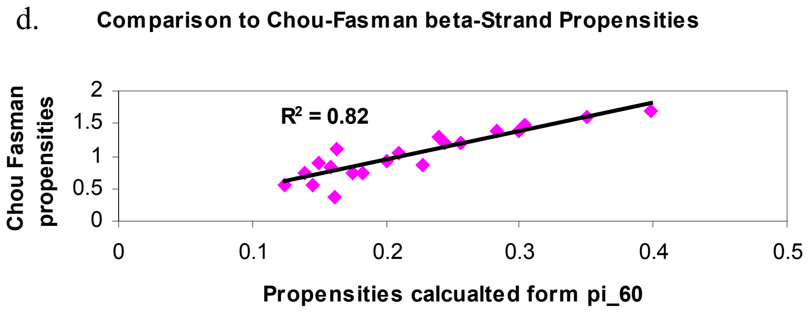
(a) 3x3 substitution matrix of secondary structure element at 80% sequence identity level (SS80). (b) 20x20 substitution matrix of amino acid at 80% sequence identity level (AA80). The red cells are the top 4 pairs with score > 1.0. The two shaded amino acids, C and P, have all negative substitution scores with all other amino acids. (c) url of the 60x60 substitution matrix of combined symbols at 80% sequence identity level (AS80), which is available for download.

Figure 5.13 Secondary Structure Propensities Extracted from 60x60 Substitution Matrix



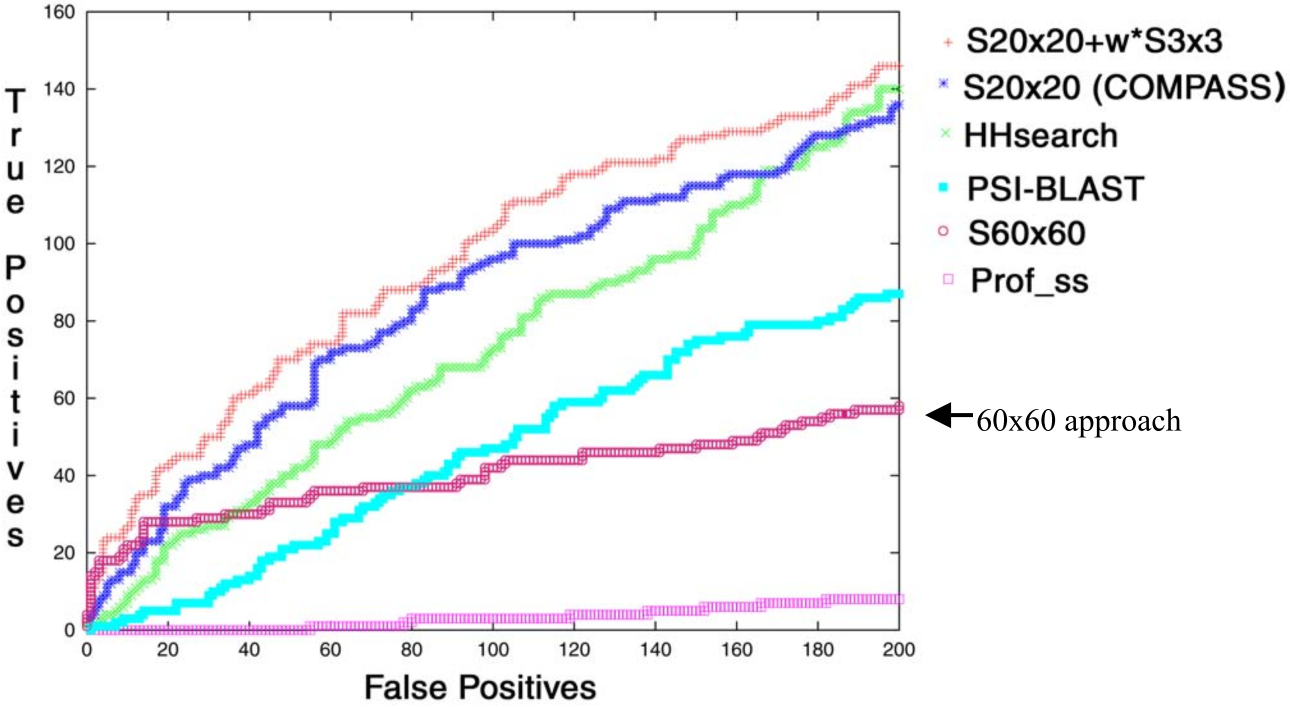






Distributions of (a) alpha-helix and (b) beta-strand propensities by amino acid types extracted from background frequencies of the 60x60 substitution matrix. Comparisons of (c) helix and (d) strand propensities we calculated with the Chou-Fasman propensities.

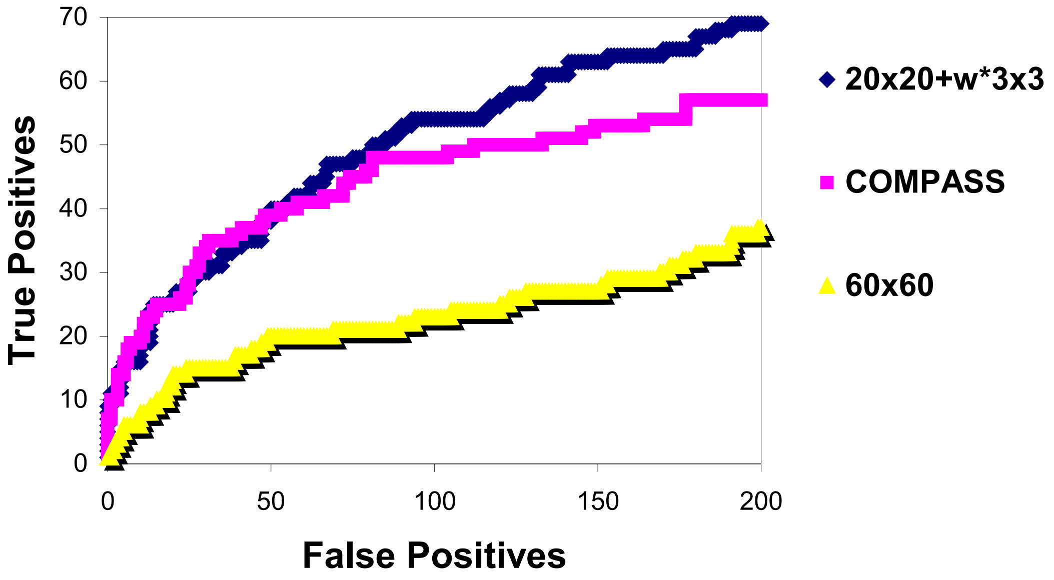
Figure 5.14 Homology Detection Ability of 60x60 Approach Fails



ROC curve analysis showing the homology detection abilities of various methods in the 200 false positives region. 60x60 approach performs worse than 20x20+3x3, COMPASS and HHsearch.

Figure 5.15 ROC Curve Analysis and Family-based Paired T-Test Result of Eight Pfam Families

a.



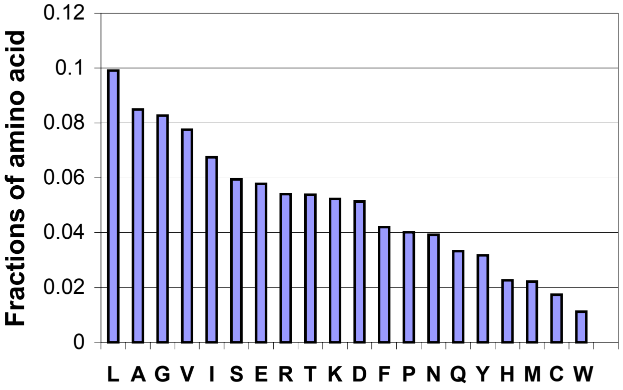
b.

	20x20+3x3	COMPASS	60x60
20x20+3x3	*		
COMPASS	+	*	
60x60	+	+	*

(a) ROC curve analysis of 20x20+3x3, COMPASS and 60x60 approaches using eight Pfam families at queries within region false positive 200. Different colors indicate different programs. (b) Family-based paired t-test results of the three approaches. + indicates a statistically significant difference at the 5% significant level. * indicates self-comparison, which is meaningless.

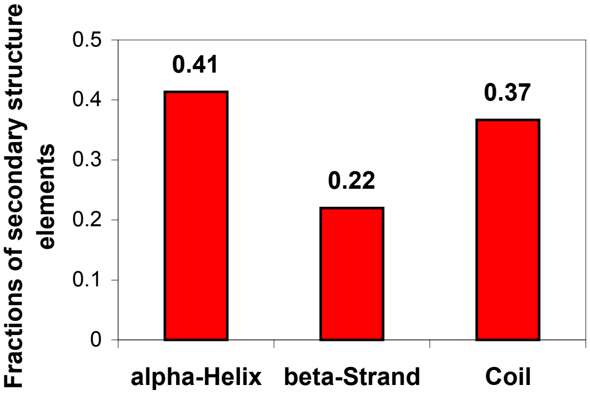
Figure 5.16 Distributions of Amino Acids and Predicted Secondary Structure Elements in Blocks+ Database

a.



(a) Distribution of amino acids in our dataset (Blocks+ 10.0). (b) Distribution of predicted secondary structure elements in our dataset. (c) Correlation analysis between our amino acid frequencies and Robinson amino acid frequencies.

b.



c.

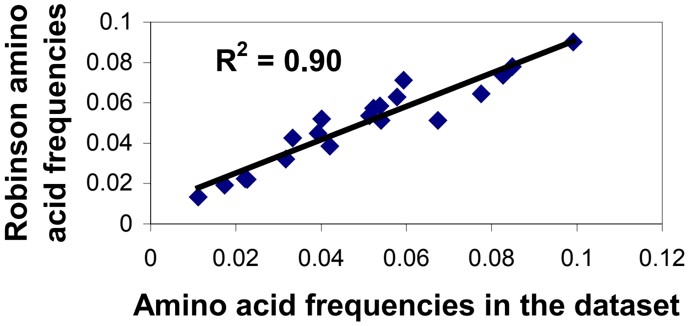


Figure 5.17 Comparison of the 60x60 Substitution Matrix and the Independent Matrix

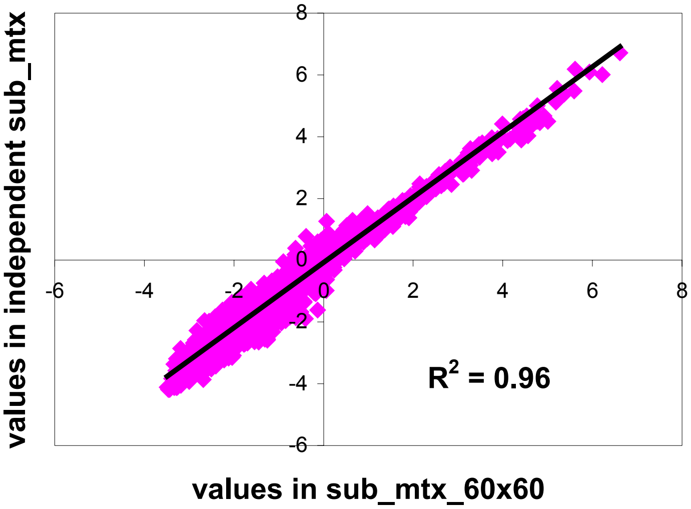
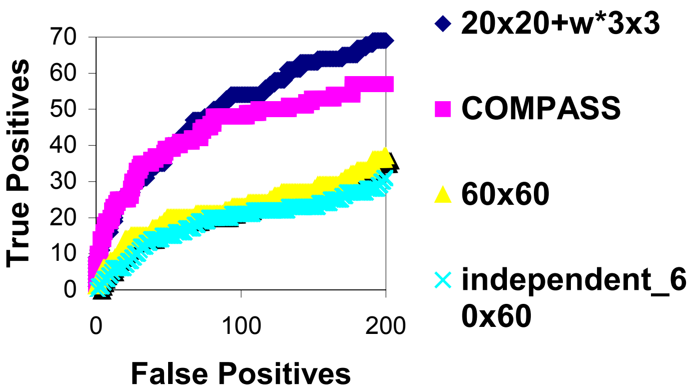


Figure 5.18 Comparison of Homology Detection Ability Effects of 60x60 Substitution Matrix and Independent Matrix

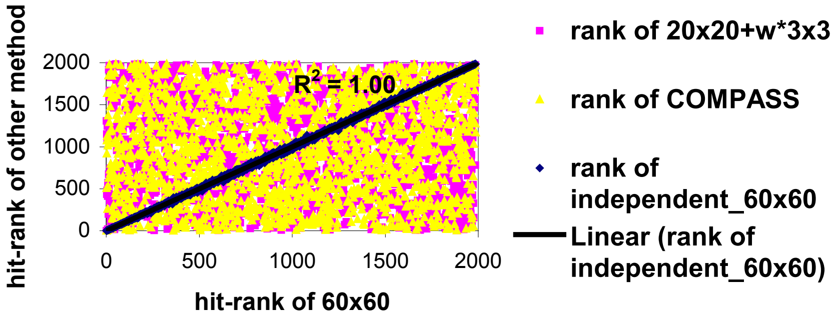
a.



b.

	60x60	Independent_60x60
60x60	*	
Independent_60x60	-	*

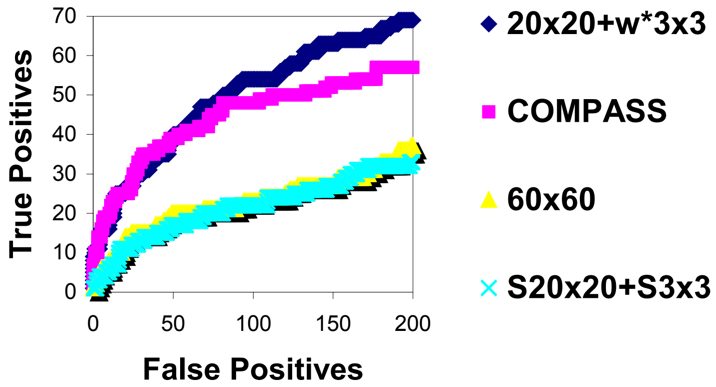
c.



(a) ROC curve analysis of 20x20+3x3 approach, COMPASS, 60x60 approach and independent 60x60 substitution matrix approach. The ROC curves of 60x60 and independent 60x60 almost overlap with each other. Different colors indicate different programs. (b) Family-based paired t-test between 60x60 and independent 60x60 shows no statistically significant difference at 5% significant level. (c) Hit-rank comparison between 60x60 and 20x20+3x3, COMPASS show no correlation at all, but there is an excellent correlation (linear regression determinant R^2 =1.00) between the hit-ranks of the 60x60 approach and the independent 60x60 matrix approach.

Figure 5.19 Comparison of Homology Detection Ability Effects of 60x60 Scoring System and 20x20+3x3 Scoring System with Equal Weights 1

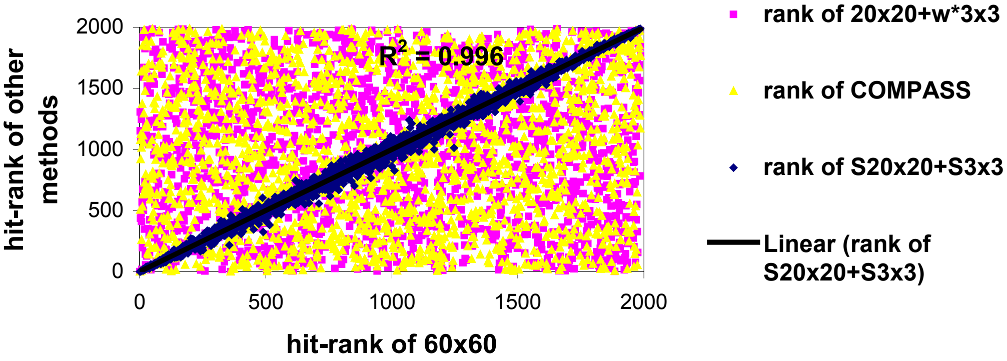
a.



b.

	60x60	S20x20+S3x3
60x60	*	
S20x20+S3x3	-	*

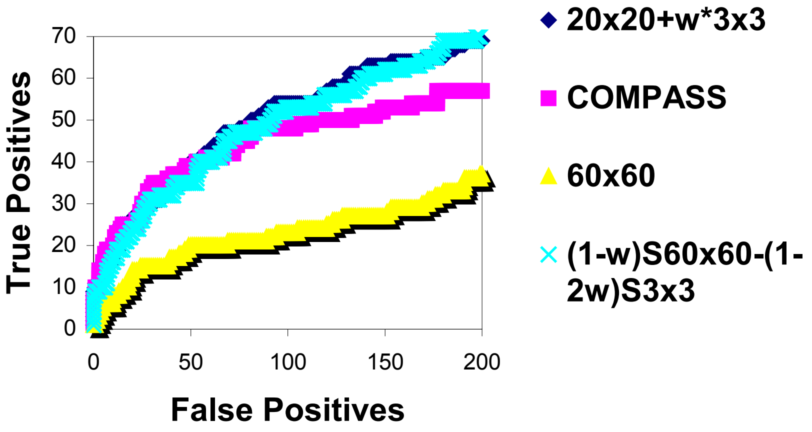
c.



(a) ROC curve analysis. The curve of 60x60 and S20x20+S3x3 approaches overlap with each other. Different colors indicate different programs. (b) Family-based paired t-test shows no statistically significant difference between the 60x60 and S20x20+S3x3 approaches at 5% significant level. (c) The hit-rank comparison shows an excellent correlation (linear regression determinant R^2 =0.996) between the ranks of 60x60 and S20x20+S3x3 approaches.

Figure 5.20 Comparison of Homology Detection Ability Effects of Weighted 60x60 Scoring System and the 20x20+3x3 Scoring System

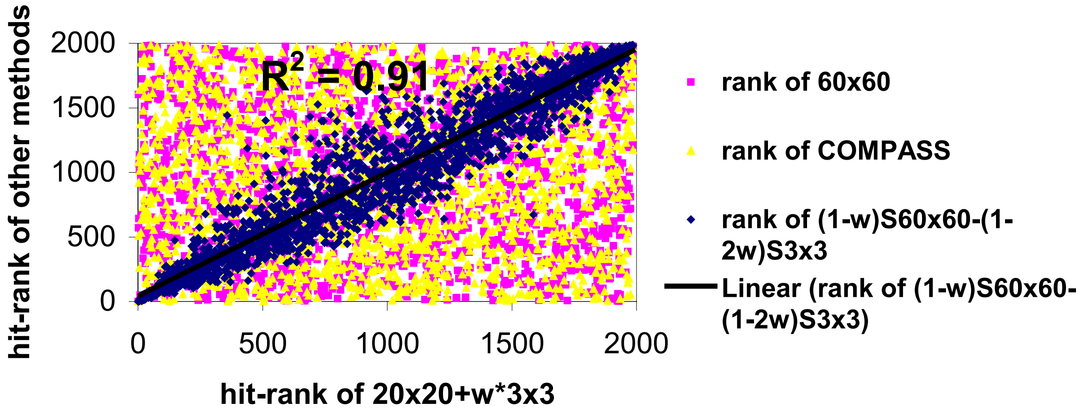
a.



b.

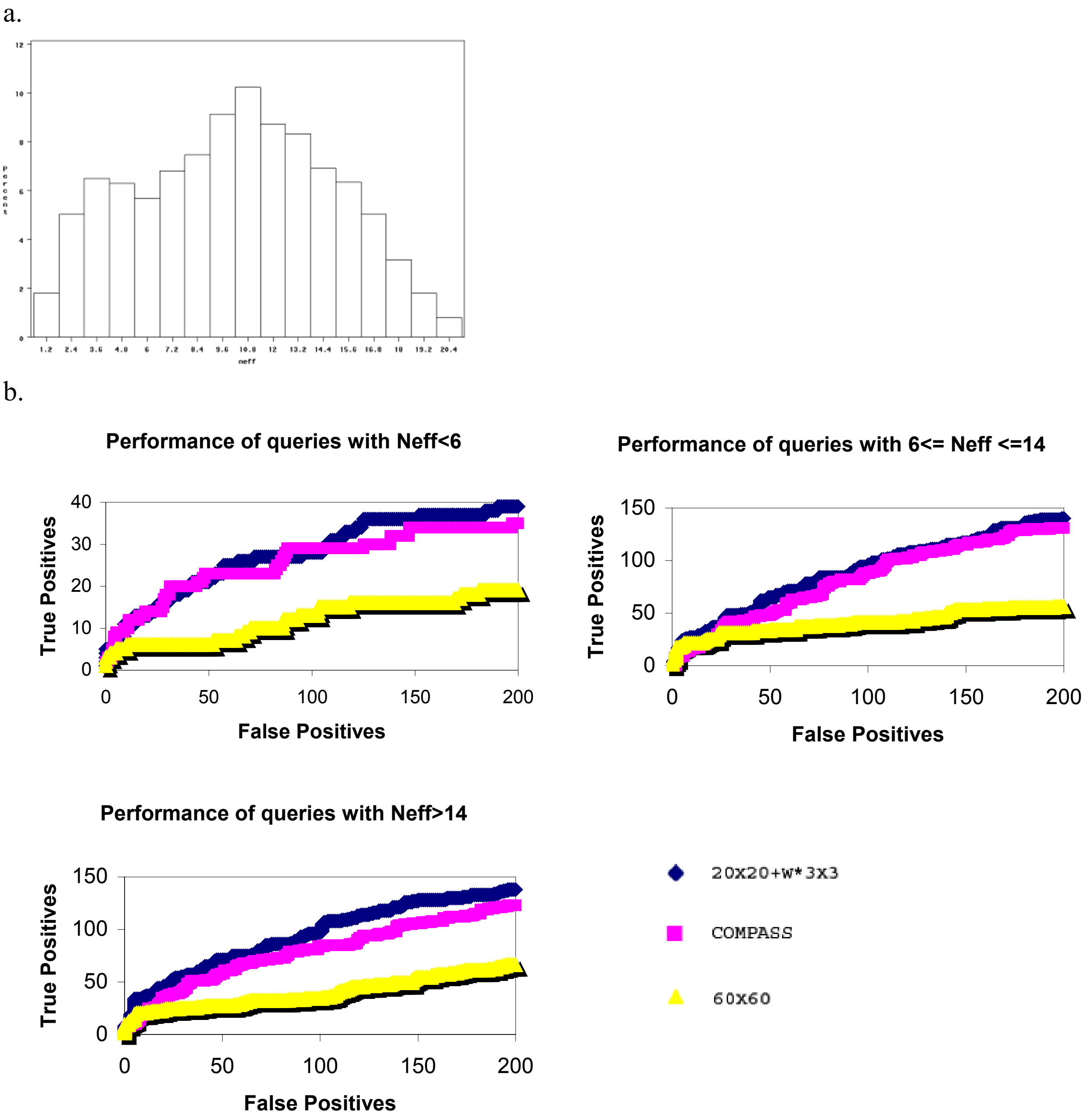
	20x20+3x3	(1-w)S60x60-(1-2w)S3x3
20x20+3x3	*	
(1-w)S60x60-(1-2w)S3x3	-	*

c.



(a) ROC curve analysis. The curves of the 20x20+3x3 approach and the weighted 60x60 approach ((1-w)S60x60-(1-2w)S3x3) overlaps. Different colors indicate different programs. (b) Family-based paired t-test shows no statistically significant difference between the 20x20+3x3 approach and the weighted 60x60 approach at 5% significant level. (c) The hit-rank comparison shows a strong correlation (linear regression determinant R^2 =0.91) between the ranks of the 20x20+3x3 approach and the weighted 60x60 approach.

Figure 5.21 Effects of the Number of Effective Count



(a) Distribution of number of effective count (Neff) of the Pfam families. (b) ROC curve analysis of queries in three categories: Neff < 6, 6-14, and Neff > 14. In all three categories, the 60x60 performs the worst and the relative performance of the three methods (20x20+3x3, COMPASS, 60x60) does not differ.

CHAPTER 6:
A Comprehensive System to Evaluate Structure Modeling Ability of Sequence Similarity Search Methods

6.1 INTRODUCTION

6.1.1 Background

Proteins with known 3-dimensional structures can serve as structure templates for homologs with unknown structures. Protein sequence similarity search and alignment methods have been used for structure modeling purposes. New programs come into being all the time and are still under development (Yona and Levitt 2002; Sadreyev and Grishin 2003; Chung and Yona 2004; Ginalski, von Grotthuss et al. 2004; Soding 2005)(Chapter 5). Therefore, it is important to have an evaluation system that provides a platform to compare their performance in terms of structure modeling abilities. Such an evaluation system helps us to understand the achievements and limitations of the field and helps researchers to choose the appropriate programs to use for their purposes.

There are two community-wide assessments for protein structure predictions, CASP (Critical Assessment of Techniques for Protein Structure Prediction) and CAFASP(Critical Assessment of Fully Automated Structure Prediction). CASP (Moult, Fidelis et al. 2005) aims at assessing manual/semi-automatic predictions, either all manual or program-generated results with human intervention. The evaluation procedures and measures that are used in CASP largely depend on the assessors and change from meeting to meeting. Many steps of the assessment, for instance, domain definition and target classification, require expert knowledge and visual inspections. Besides, it is limited to a small number of test proteins (63 For CASP6). CAFASP (Fischer, Rychlewski et al. 2003) aims at assessing automatic structure prediction servers. Unlike CASP, CAFASP uses automatic evaluation programs. However, CAFAST is also limited to a small number of testing proteins. LiveBench

(Rychlewski and Fischer 2005) is an evaluation server that also evaluates the performance of automatic structure prediction servers. It is an extension and complement to CAFASP by continuously assesses a relatively large number of predictions every week. Another large-scale evaluation project for automatic structure prediction servers is EVA (Eyrich, Marti-Renom et al. 2001), which together with LiveBench provides complements to CAFASP.

6.1.2 Objective

Since there is no standard way to perform the evaluation, we decided to develop an automatic large-scale evaluation system to evaluate different aspects of the structure modeling ability of sequence similarity search programs. In order to set up a systematic and comprehensive evaluation system, we first need to select a representative testing dataset that is non-biased and is of certain degree of difficulties, and then need to combine different sequence and structure similarity measures, as well as measures from CASP and LiveBench, to assess the sensitivity and specificity of different programs. The evaluation procedure needs to include assessment for both fold recognition abilities and alignment qualities from global and local perspectives using both reference-dependent and reference-independent approaches.

6.2 EVALUATION ALGORITHM DEVELOPMENT

6.2.1 Selection of Representative Dataset

Based on the criteria for the representative dataset, non-biased and of certain degree of difficulties, we decided to select a dataset with a maximum ~20% pairwise identity, which is within the twilight zone, from SCOP (Murzin, Brenner et al. 1995) domain sequences. Astral (Brenner, Koehl et al. 2000; Chandonia, Hon et al. 2004) offers a SCOP domain sequence dataset of 40% identity (SCOP40 set) that contains good quality structures. We select our 20% dataset out of the SCOP40 set. Astral also offers SCOP domain set of maximum 20% identity (SCOP20). However, since the Astral SCOP20 set is derived based on sequence alignment method (BLAST), and sequence-based alignments are not accurate at this level of sequence similarity, we decide not to use it but to select our 20% dataset based on alignments generated by structure-based methods. Three structure-based alignment methods, DALI (Holm and Sander 1995), TM (Zhang and Skolnick 2004) and FAST (Zhu and Weng 2005), were chosen because they are either known to generate good quality alignments or are fast to run.

Methods for pairwise identity calculation

Three means for calculation of pairwise sequence identities are used: (1) percentage identity within the aligned blocks, $pid(1) = \frac{N_{id}}{L_{ali}}$, where N_{id} is the number of aligned identical residues, L_{ali} is the length of the aligned regions; (2) percentage identity over the shorter query, $pid(2) = \frac{N_{id}}{L_{shorter}}$, where $L_{shorter}$ is the length of the shorter query of the compared pair; (3) real identity within the aligned blocks combined with a random identity within the unaligned regions, $pid(3) = \frac{N_{id} + L_{unali} * pid_{random}}{L_{shorter}}$, where L_{unali} is the length of the unaligned regions of the shorter query, pid_{random} is the percentage identity of random alignments. Estimation of the random identity as $\sum_{i=1}^{20} f_i^2$, where f_i is the frequency of amino acid i in SCOP40 set, results in 6%. Using Dayhoff amino acid frequencies (Dayhoff 1978) also results in 6%. From experience, the random identity is in the range of 5-8%. Screening of this region gives similar results and thus 6% is used in the final calculation. Variations of these three means of identity calculation (Appendix A.1) are also tried and give similar results.

Finding the largest number of unique representatives for each superfamily

To ensure each superfamily in SCOP has at least one representative in the dataset, the representative selection process uses each superfamily as a unit. We use a novel method to select representatives for each SCOP superfamily. The criterion is to find the largest number of domains in this superfamily that do not have a pairwise identity higher than the cutoff. For one SCOP superfamily, we first calculate pairwise identities for every pair of sequences in this superfamily in the SCOP40 set. Using a graphic representation, each domain sequence is a vertex; if the pairwise identity between two sequences is above certain cutoff x%, an edge is placed between them (Figure 6.1). A dynamic programming-based method is then employed to find the largest number of domain sequences that do not have an edge between each other.

Selection of the entire representative dataset

The entire representative set of cutoff x% is composed of these domain sequences. This process is done for DALI-, TM- and FAST-generated structural alignments. Because there are intrinsic differences between different structure alignment programs, we want to utilize an identity measure that best reflects the evolutionary distance between protein pairs in spite of different structure alignment programs. Thus the best identity measure should pick the largest overlapping representatives between all three programs and leave the least number of unique representatives for each program. By screening the identity cutoff x% from 15-25%, all three means of identity calculation result in ~3500-4500 representatives. When fixing the size of the overlapping representatives to approximate 4000, we can see from Figure 6.2 that identity measure *pid*(1) is clearly the worst one, and identity measure *pid*(3) is able to choose a similar number of representatives for each method and has the least average unique representatives (2.7%). Thus, the final dataset is chosen using identity measure *pid*(3), and includes the OR set of representatives from all three structure alignment methods.

6.2.2 Reference-Dependent Evaluation Of Structure Template Quality

For reference-dependent evaluation, in order to assess whether the hit is a good global structure template for the query, we consider the structure and sequence similarity scores of the reference alignment (structure-based alignment) only. In this step (Step 1), the 3d-strucures of the query and hit are optimally superimposed onto each other using least-square minimization based on the reference alignment. All structure scores are then calculated from this structure-based superposition of the pair.

SCOP superfamily/fold/class levels could serve as a standard for true/false to calibrate these scores. If we are lucky, the distributions of these scores for domain pairs within the same superfamily and between different superfamilies or different folds should separate very well, allowing for an easy discrimination of a good structure template from a bad one. Unfortunately, this is not the case. Each of the individual score provides a poor separation even between the same superfamily and different classes (data not shown). Therefore, we decide to use the Support Vector Machine (SVM) technique (Joachims 1999) to combine all the scores in a reasonable way to distinguishing good and bad templates.

Selection of SVM features based on SCOP classification

Eight types of scores are used initially, including sequence scores such as identity, blosum score, coverage, and structure scores such as GDT_TS (Zemla 2003), match index (Kolodny, Koehl et al. 2005), DALI Z-score (Holm and Sander 1998), TM score (Zhang and Skolnick 2004) and FAST score (Zhu and Weng 2005). Each type of score is calculated for all three types of reference alignments (DALI/FAST/TM) of all pair of domains in the testing dataset. SCOP classification is used as a reference for true/false in SVM training. Because we know that many different SCOP superfamilies are homologous to each other, and many different SCOP folds are actually of the same structural fold (e.g. Rossmann fold domains), to avoid ambiguity, we select two stringent classification levels (superfamily and class) as standard. In our SVM training, domain pairs belong to the same superfamily are considered true, different classes false. 2000 pairs of SCOP domain are randomly chosen for SVM linear model training and the initial resulting classification accuracy is 94.8%. In this testing, totally 30 features (combinations of scores and types of structure alignments) are used as SVM inputs. We find out that removing some input scores would result in better classification accuracy. In order to find out the importance of the individual features, we calculate and compare the standardized weights of the linear model by normalizing the scores by mean and standard deviation. Five parameters are found to dominate the classification effect and give the best prediction accuracy (95.7%) (Figure 6.3a). According to their rank of importance, these include DALI Z-score (native), Fast score (native), GDTTS of TM alignment, coverage of FAST alignment, and blosum score of DALI alignment. The calculated DALI Z-score of FAST and TM alignments also have large weights, but since they highly correlate with native DALI Z-score ($R^2 = 0.95$), adding them in do not increase the prediction accuracy. Thus, to avoid redundancy, we do not put them in the final parameter list.

Selection of SVM score cutoffs to allow for unknowns

Since relationships between some structural folds are unclear at present (for instance, Rossmann fold and TIM barrel fold), it is problematic to judge if hits between these kinds of folds are true or false. Therefore, we decide to select two SVM score cutoffs in order to allow for unknowns besides true or false hits. In order to take into account both the expert-curated homology/fold-similarity relationship and the well-established automatic structure similarity estimation methods, we use both SCOP classification and SVM score cutoffs as criterion for true/false/unknown. If a hit belongs to the same SCOP superfamily as the query, we consider it as true. Otherwise, hits with scores higher than the high-cutoff of SVM score is considered true, lower than the low-cutoff of SVM score false, and in between unknown.

To decide on the high-cutoff, four representative problematic fold pairs for each of the four major structure classes are selected and their SVM score distributions are plotted (Figure 6.3b, Table 6.1a). Since we do not want to include these problematical fold pairs as true hits, the 95% percentile of their distributions are calculated and their average is taken as the high-cutoff (Table 1). On the other hand, some protein domains belonging to the same structural folds may not have high structural scores to each other because of many insertions and deletions (for example, Rossmann fold domains), but we do not want these domain pairs to be judged as false hits. Therefore, four representative such structural folds are selected for each of the major class, and the average of the 5% percentiles of their SVM score distributions is taken as the low-cutoff (Figure 6.3c, Table 6.1b).

6.2.3 Reference-Dependent Evaluation Of Alignment Quality

For reference-dependent evaluation, in order to access the quality of a sequence-based alignment between query and hit in terms of its usefulness for structural modeling, we compare the sequence-based alignment to the structure-based reference alignment of the pair. Structure alignments generated by DALI are used as reference alignments in the following studies for DALI is known to generate good quality alignments. In this step (Step 2), the 3d-strucures of the query and hit are optimally superimposed onto each other according to the sequence-based alignment. All scores are calculated from this sequence-based superposition of the pair.

Two type of structure modeling scores, GDT_TS (Zemla 2003) and LiveBench 3dscore, have been used traditionally in CASP and other assessments (Ginalski, Grishin et al. 2005; Rychlewski and Fischer 2005) to evaluate the quality of a sequence alignment. However, these scores have only been used to rank different structure models, while no cutoff has ever been given for a decent alignment. Another type of score, number of correct matches, has also been use to access alignment quality (Sadreyev and Grishin 2003). Number of correct matches is the number of residue pairs that are aligned the same way in sequence alignment as in the structure alignment. In order to find a reasonable cutoff for alignment quality, we decide to use all three types of scores. However, during our tests, GDT_TS and LiveBench 3dscore give very similar results and conclusions (Figure 6.4 a & c). As GDT_TS is a more popular measure, we use GDT_TS and number of correct matches, not LiveBench 3dscore, in our criteria for alignment quality.

To calibrate GDT_TS for a decent alignment, we randomly chose 500 domains and generate pairwise sequence alignments in an all-against-all fashion. We then apply test step1 to these alignments and use alignments of the false hits from step 1 as negative controls and PSI-BLAST alignments with significant E-values (less than default E-value cutoff: 0.005) as positive controls, and mark them on the 2D plot of GDT_TS of structure alignment vs. sequence alignment (Figure 6.4a). These significant PSI-BLAST alignments are all true hits from step 1. From the distribution of GDT_TS of the false hit alignments, the 95th and 99th percentiles are taken as the potential cutoffs. However, since there is no clear separation between the positive controls and negative controls, there are alignments with significant E-values have GDT_TS less than the potential cutoffs. In order to include these hits as good ones, we use number of correct matches to calibrate. Alignments of the true hits from step 1 are compared with structure alignments (DALI) and the number of correct matches is calculated for each alignment. These alignments are then mapped to the PSI-BLAST alignments with significant E-values on the 2D plot (Figure 6.4b). Previous studies (Sadreyev and Grishin 2003) show that number of correct matches 5 is a reasonable cutoff for alignment comparisons. From the mapping in Figure 6.4b we know that alignments with number of correct matches more than 5 covers about 97% of all PSI-BLAST alignments with significant E-values. Thus 5 is chosen to the cutoff for number of correct matches.

6.2.4 Summary Of Reference-Dependent Evaluation Criteria

For reference-dependent evaluation, our criteria for global structural template quality are shown in Figure 6.5a. If the hit and query belong to the same SCOP superfamily or their SVM score is higher than the 0.6, the hit is considered true; if the hit and query do not belong to the same SCOP superfamily and the SVM score is lower than the –0.6, it is considered false; if the hit and query do not belong to the same SCOP superfamily and their SVM score is between –0.6 and 0.6, it is considered unknown. This way, we take into account both the expert-curated homology/fold-similarity relationships (SCOP superfamily) and the combinations of well-established automatic methods to estimate the structure similarities.

Our criteria for sequence alignment quality are shown in Figure 6.5b. If the sequence alignment has a GDT_TS higher than 0.15 or has a number of correct matches more than 5, it is considered true.

6.2.5 Reference-Independent Global Mode Evaluation

Since the reference-independent evaluation is based on sequence alignment only, we can just use GDT_TS of the sequence alignment to evaluate fold similarity and alignment quality over the entire length of the query domains. The reference-dependent evaluation studied above is a global mode evaluation and is done in two steps, where GDT_TS is used in the second step. When figuring out the cutoff for GDT_TS, we already take into consideration the true and false hits from step 1. And thus 99% of the hits that do not share fold similarity with the query are excluded by the GDT_TS cutoff. Therefore, we can simply take the GDT_TS cutoff (>= 0.15) figured out in the reference-dependent evaluation and apply it to the global mode of reference-independent evaluation.

6.2.6 Reference-Independent Local Mode Evaluation

For reference-independent local mode evaluation, we cannot use local GDT_TS (lGDT_TS) directly because it has a significant dependency on aligned domain length. The formula of local GDT_TS is as follows,

Equation 6.1

$$lGDT\_TS = \frac{1}{L_{ali}} \sum_{L_{ali}} \left(\frac{n_1 + n_2 + n_4 + n_8}{4} \right)$$

where n1, n2, n4, n8 are the numbers of aligned C-Alpha atoms that are within the distance of 1, 2, 4, 8 Å from each other, and L_{ali} is the length (number of residues) of the aligned region. From Equation 6.1 we can see that if the aligned region is very short, essentially all residues in the aligned region will be very close to each other and the resulting lGDT_TS will be artificially large. In the extreme case, if there is only one residue aligned, the lGDT_TS will be a perfect score of 1.0. The length dependency effect also exists in the global

GDT_TS, but it does not affect the global mode evaluation. When the aligned region is very short compared to the length of the query, the resulting global GDT_TS is small, too (refer to Equation 6.3). Thus we can see that the global GDT_TS favors long alignments while the local GDT_TS favors short alignments.

In order to eliminate this length effect for local GDT_TS, we first model the length-dependency of lGDT_TS, and then normalize the local GDT_TS scores by the model.

Modeling of the length-dependency of local GDT_TS

For a particular length L, we randomly select 1000 pairs of domain fragments of length L from our testing dataset. Each pair of fragments are forced to align with each other from end to end and optimally superimposed to each other according to this alignment. lGDT_TS score is calculated based on this superposition. Thus, for length L we have 1000 values of random lGDT_TS scores and their mean and standard deviation (sd) are calculated. By repeating this process for lengths 3 to 500, we are able to plot the length-dependency of lGDT_TS (Figure 6.6a). When transforming both x- and y-axes to log-scale, the scatter plots of the mean and sd of lGDT_TS show a linear trend with respect to length (Figure 6.6b), which indicates a power law relationship between the values and length. Using power law function $f(L)=cL^b$ to fit the lGDT_TS mean and sd (Figure 6.6b), we get the length-dependency models of lGDT_TS mean and sd:

Equation 6.2 $mean(L) = 3.807L^{-0.956}$, $sd(L) = 0.617L^{-0.714}$.

(The linear model parameters are fitted using SAS STAT package.)

Normalization of local GDT_TS scores

The raw lGDT_TS score of the sequence alignment is normalized to a Z-score by the transformation $Zscore = \frac{raw\_lGDT\_TS - mean(L)}{sd(L)}$, where L is the length of the aligned region in the sequence alignment. By using the Z-score of lGDT_TS, we are able to screen for alignments with lGDT_TS scores significantly higher than random scores and thus are true ones (i.e. local alignments of good quality). In order to obtain a reasonable Z-score cutoff for true alignment, we initially choose Z-scores 3 and 5 as cutoff candidates and compare their corresponding local GDT_TS values to those of the 95 percentile of local GDT_TS distribution (Figure 6.7a). The comparisons are made for lGDT_TS distributions of alignment lengths 5, 20, 50, 100, 200 and 500. From Figure 6.7a we can see that both Z-scores 3 and 5 cutoff values are more stringent than the values of 95 percentile, but Z-score 3 values are more similar to those of 95 percentiles. In an example distribution of lGDT_TS of alignment length 50 (Figure 6.7b), the value of Z-score 3 cutoff is more extreme to that of the 95 percentile. Thus Z-score 3 is decided to be the cutoff for true good alignments.

6.2.7 Summary Of Reference-Independent Evaluation Criteria

For reference-independent evaluation, we use GDT_TS score derived from sequence-based alignment as criteria for overall fold similarity and alignment quality evaluation, but there are differences between global and local modes evaluations. To evaluate global mode, we use GDT_TS directly (Figure 6.8a). If the GDT_TS is higher than 0.15, the hit is considered true. To evaluation local mode, we first transform the local GDT_TS score of an alignment into Z-score, and then judge true or false according to the Z-score (Figure 6.8b). If Z-score is higher than 3, this alignment is considered true.

6.3 RESULTS AND DISCUSSIONS OF EVALUATIONS OF SEQUENCE SIMILARITY SEARCH PROGRAMS

This comprehensive evaluation system enables us to compare the performances of different sequence alignment methods. By setting up the cutoffs for different steps and modes of structure modeling efficiency (Figure 6.5, Figure 6.8), we are able to judge if a hit generated by a sequence alignment method is true positive (TP) or false positive (FP) according to different modeling purposes, which in turn enables us to use ROC curve, a sensitivity and specificity evaluation technique, to compare the performances of different programs.

For each method to be evaluated, the hits need to be sorted by their E-values in an ascending order. A ROC curve is then generated by plotting the numbers of true positives corresponding to each increment in the number of false positives. In the ideal case, a method should find all the true positives before finding any false positives, and the curve should go vertically up from zero and then horizontally right. Thus, the further top-left the curve goes, the better the method is.

We use our evaluation system to compare the performances of six selected sequence similarity search programs, including the popular method PSI-BLAST (Altschul, Madden et al. 1997), profile based methods COMPASS (Sadreyev and Grishin 2003) and HHsearch (Soding 2005), combined profile and secondary structure methods PROF_SS (Chung and Yona 2004), COMPRASS (Chapter 5) and HHsearch_ss (Soding 2005). These programs are run on the entire representative dataset in an all-against-all fashion and generate six sets of sequence-based alignments. The evaluation methods are applied to these sequence alignments and the required scores are calculated. ROC curves are then plotted to compare the programs.

6.3.1 Evaluations On A Small Testing Set

For the purpose of getting a testing result quickly, a small testing set with 500 domain sequences is chosen to perform the evaluation upon. The testing set domains are selected from the 4147 representative dataset. To ensure there is enough number of homologs (true positives) in the testing set, these domains are chosen in a special way. One domain (head domain) is chosen randomly from the entire dataset first, we then find all other domains in the dataset that belongs to the same SCOP superfamily as the head domain and add them in the testing set. This process is repeated until the number of domains in the testing set reaches 500. The performance comparison results shown in this section are based on this testing set.

Reference-dependent structure template quality evaluation is carried out first. For alignments generated by each program, the hits are sorted by their E-values (PSI-BLAST, COMPASS, COMPRASS) or p-values (Prof_ss, HHsearch, HHsearch_ss). The TP and FP are decided according to our criteria shown in Figure 6.5a and a ROC curve is generated for each method. Figure 6.10 shows the resulting ROC curves. From bottom-up, we can see that PSI-BLAST performs worst. The performances of COMPRASS and COMPASS are comparable but COMPRASS is slightly worse than COMPASS. HHsearch and Prof_ss are comparable and HHsearch_ss performs the best.

For reference-dependent alignment quality evaluation, the TP and FP for each method are decided according to the criteria shown in Figure 6.5b. Figure 6.11 shows the resulting ROC curves. From bottom-up, the performance increases with programs PSI-BLAST, Prof_ss, COMPASS, HHsearch, COMPRASS and HHsearch_ss. This ranking of program performances is consistent with the results obtained by the HHsearch author (except COMPRASS, which was not available to the HHsearch author). The increase in performance for COMPRASS compared to COMPASS is similar to that of HHsearch_ss compared to HHsearch.

For reference-independent global mode evaluations, the TP and FP for each method are decided according to the criteria shown in Figure 6.8a. Figure 6.12 shows the testing result. From bottom-up, the performance increases with programs PSI-BLAST, COMASS, COMPRASS and HHsearch (the two are comparable to each other), Prof_ss, HHsearch_ss. The degree of improvement of COMPRASS over COMPASS is very similar to that of HHsearch_ss over HHsearch.

For reference-independent local mode quality evaluation, the test is done according to the TP/FP criteria in Figure 6.8b. The resulting ROC curves are shown in Figure 6.13. According to this figure, the ability of detecting local structure similarities increases with programs PSI-BLAST, COMPASS, COMPRASS and HHsearch (comparable), HHsearch_ss, Prof_ss.

HHsearch_ss program performs the best in all categories except for local mode of reference-independent evaluation. PROF_SS performs best in the local mode evaluation but not in the global mode evaluations, which indicates that it generates mostly locally optimal sequence alignment. The fact that PSI-BLAST performs the worst in every categories indicates that it do not perform well on dataset of this degrees of difficulties.

Except for the overall template detection, COMPRASS performs better than COMPASS, and the increase in performance is comparable to the increase of HHsearch_ss over HHsearch, which indicate that adding predicted secondary structure information to profile information indeed helps to increase the sequence alignment quality. And although the secondary structure information is incorporated in different ways, the amounts of information added in are the same, and thus resulting in the same amount of effects.

6.3.2 Evaluations On The Entire Representative Set

The same evaluation procedure is done on the entire 4147 representative dataset. Testing results are shown in Figure 6.14. For reference-dependent structure template quality evaluation, the resulting ROC curves are shown in Figure 6.14a. From bottom-up, the programs with performances ranking from worst to best are PSI-BLAST, HHsearch, Prof_ss, COMPASS, COMPRASS, HHsearch_ss. The performance increase of COMPRASS over COMPASS is smaller than that of HHsearch_ss over HHsearch. For reference-dependent alignment quality evaluation, the resulting ROC curves are shown in Figure 6.14b. The programs with performances ranking from worst to best are PSI-BLAST, Prof_ss and COMPASS and HHsearch (the performances of the three programs are comparable), HHsearch_ss, COMPRASS. The performance increase of COMPRASS over COMPASS is larger than that of HHsearch_ss over HHsearch. For reference-independent global mode evaluation, the resulting ROC curves are shown in Figure 6.14c. The programs with performances ranking from worst to best are PSI-BLAST, COMPASS, COMPRASS, Prof_ss, HHsearch, HHsearch_ss. The degree of performance increase of COMPRASS over COMPASS is very similar to that of HHsearch_ss over HHsearch. For reference-independent local mode quality evaluation, the resulting curves are shown in Figure 6.14d. The programs with performances ranking from worst to best are PSI-BLAST, COMPRASS and HHsearch (the performances of the two are comparable), HHsearch_ss, COMPASS, Prof_ss.

Overall, the performance ranks of programs on the entire dataset are similar to those on the small testing set shown in the above section, which indicates that our evaluation system is robust. However, there are two major differences comparing the results on the two datasets. One difference occurs for reference-dependent structure template evaluation (Figure 6.14a). On the small testing set, COMPRASS performs slightly worse than COMPASS, while on the entire set, COMPRASS performs much better than COMPASS. The other difference occurs for reference-independent local mode evaluation (Figure 6.14d). COMPRASS performs better than COMPASS on the small set, but worse on the entire set. The performance comparison between COMPRASS and COMPASS on the entire representative dataset makes more sense. Since COMPRASS tends to generate more global alignments (i.e. long alignments with large coverage), while COMPASS alignments tend to be more local (i.e. short alignments), it is reasonable for COMPASS to perform better on the local mode but worse on the global mode. As to the reason why COMPASS performs better than COMPRASS for reference-dependent structure template evaluation (global mode) on the small set, it is probably because the small set is more compact with close homologs that are more easily detected by COMPASS, while COMPRASS tends to detect more remote homologs.

Looking at all four evaluation results on the entire dataset, PSI-BLAST performs worst in every category, which indicates that PSI-BLAST does not perform well on dataset of this degrees of difficulties (within and below the twilight zone). Except for the reference-independent local mode evaluation, COMPRASS performs better than COMPASS, and the degree in performance increase is comparable to that of HHsearch_ss over HHsearch. This observation indicates that adding predicted secondary structure information to profile information indeed helps to increase the sequence alignment quality.

Comparing the global performance of HHsearch_ss and COMPRASS clearly shows that HHsearch_ss is better at detecting overall structure template (Figure 6.14a) while

COMPRASS is better at generate correct alignement (Figure 6.14b). And when testing for combined global structure template detection ability and alignment quality (i.e. the reference-independent global mode, Figure 6.14c), COMPRASS performances worse than HHsearch_ss. These results give comprehensible indications of the advantage and limitations of the programs. If parts of the query and template structures that are aligned are dissimilar to each other, the overall superposition of the two domains could be skewed so that the correctly aligned equivalent residue pairs are distant from each other. Since GDT_TS measure the distance between equivalent residue pairs, when parts of the aligned structures have low structural similarity, even if the alignment between them is correct, the overall GDT_TS score could still be low. Since the combined global structure similarity and alignment quality (i.e. the reference-independent global mode) is measured by GDT_TS, the method that has poorer structure template detection ability (COMPRASS) would have a lower score than the method that has better structure template detection ability but poorer alignment quality (HHsearch_ss). Thus, comparisons between the evaluation results of different criteria inform us that the method COMPRASS generates better sequence alignment but detects poorer structure template than HHsearch_ss. Another study (Pei and Grishin submitted) using similar scoring function as COMPRASS also shows that the scoring function used in COMPRASS helps increase sequence alignment quality. The reason why COMPRASS detects poorer structure template than HHsearch_ss might be that the statistics (E-value calculation method) used by COMPRASS are worse, for COMPRASS statistics are fitted for mixtures of different protein families, while HHsearch_ss statistics are calculated specifically for each individual families. Therefore, the ranking of hits generated by COMPRASS does not reflect the structure similarity as well as that generated by HHsearch_ss.

6.4 RESULTS AND DISCUSSIONS OF THE EVALUATION SYSTEM

SCOP is a popular protein structure classification database constructed with both structural and evolutionary considerations (section 1.1.4). Because it is mainly expert manually curated, SCOP classification is often used as a gold standard for homology relationship and structure fold similarity (Chung and Yona 2004; Zhu and Weng 2005; Paccanaro, Casbon et al. 2006). However, there are known problems with SCOP. One problem is that proteins belonging to different SCOP superfamilies could be homologous to each other. For example, thiamin phosphate synthase and Indole-3-glycerophosphate synthase (IPGS) are homologous to each other (Nagano, Orengo et al. 2002; Cheek, Qi et al. 2004), but they are assigned to two different superfamilies in SCOP (thiamin phosphate synthase and ribulose-phosphate binding barrel). Another problem with SCOP is that proteins belonging to different SCOP folds could have the same structure fold. The most obvious example is Rossmann-like fold proteins (Anantharaman and Aravind 2006). In the current version (version 1.69) of SCOP, there are 136 SCOP folds in the α/β class, while at least 77 of them are of Rossmann-like structural fold. For example, proteins in SCOP fold nucleotide-binding domain and SCOP fold FAD/NAD(P)-binding domain are all of Rossmann-like fold.

Because SCOP has these problems, it is not a good approach to use SCOP classification as gold standard blindly. Researchers have realized this problem and uses supplemental methods in addition to SCOP classification. For instance, Soding in his HHsearch paper (Soding 2005) uses two sets of criteria for true or false positives when evaluating homology detection abilities. In the first set, he defines true positives as pairs from the same SCOP superfamily, false positives as pairs from different SCOP classes. All the other pairs are considered to be unknown and are ignored. This leaves a large portion of domain pairs in a gray area (~40% unknowns according to personal communications between Drs. Soding and Grishin). Because he thinks using SCOP only and ignoring the pairs in the gray area is unfair, Soeding uses MaxSub score (Siew, Elofsson et al. 2000) in addition to SCOP superfamilies as criteria for true positives in the second set (Soding 2005). Therefore, to develop a comprehensive evaluation system based on SCOP and structural and sequence similarity is highly necessary.

6.4.1 Representative Dataset

The final representative dataset contains 4147 SCOP domain sequences. Figure 6.9a shows the distribution of domain lengths. The domain lengths range from 31 to 1256 amino acid long with a median of 151 amino acids. These representative domains belong to 1516 SCOP superfamilies. Figure 6.9b shows the distribution of number of representatives per SCOP superfamily. The 4147 representatives belong to seven SCOP classes. Figure 6.9c shows the distribution of percentage of representatives per SCOP class of this dataset, which has a similar distribution as the representatives in the Astral SCOP20 set.

Multiple sequence alignments for each sequence in the representative dataset are generated using PSI-BLAST with an inclusion E-value cutoff of 10-4 for up to 2 iterations. Secondary structures for each of the representative sequence are predicted using PSIPRED (Jones 1999). Compared to real secondary structures generated by DSSP (Kabsch and Sander 1983) based on the 3-dimensional structures of the domains, the average accuracy of the predicted secondary structure is 80% for Q3 (Chandonia and Karplus 1999) and 78% for SOV (Zemla, Venclovas et al. 1999), which are of the same prediction accuracy level as reported (Bryson, McGuffin et al. 2005).

For our purposes, it is important to have a large-scale, non-biased representative testing dataset. In addition, this dataset needs to be of certain degree of difficulty for correct homology-identification and alignment. Otherwise, all programs could perform well and thus the system loses the distinguishing power.

6.4.2 Global/local mode

The global and local modes of evaluations address different goals of structure modeling and should both be considered. Global mode has a goal of getting a good overall structural template for a query, i.e. finding a hit that share structural fold similarity to the query over the entire length. Local mode has a goal of finding local, maybe short, but precise alignments to segments of a query (i.e. fragment similarity). In this respect, fold similarity is not needed; just the structural accuracy of a local alignment is evaluated. Apparently, both modes are useful for structure modeling. It is possible that different sequence alignment programs or different versions of a program will be optimal for different modes, and thus both modes should be evaluated.

6.4.3 Reference-dependent evaluation

We need to have reference-dependent and reference-independent evaluations with structure-based alignment as the reference. Reference-dependent evaluation should be done by comparing sequence-based alignment to the reference alignment (i.e. structure-based alignment). At low sequence identity level (<20%), a structure-based alignment is more reliable than a sequence-based alignment, and can be considered as the best possible alignment for a particular pair of domains. Since for one pair of domains we have only a single reference alignment, many short but structurally equally good alignments (for example, helix aligned to helix) are not considered. Thus the reference-dependent method is a good way to evaluation global mode alignments but not local mode.

The reference-dependent global mode of evaluation should consist of two steps. Step1 is to decide whether the hit can be true in principle based on a reference, namely, to judge if the hit is overall a good structure template in the case of best alignment. Step 2 is to decide whether the hit can be true in terms of usefulness for structure modeling, namely, to judge the quality of the sequence alignment. Traditionally, SCOP superfamily/fold classification was used at step1; and step 2 has been ignored. We should make them both work.

6.4.4 Reference-independent evaluation

Reference-independent evaluation should be based just on the sequence alignment itself. It does not need reference structure alignments or reference classifications. Thus it is more flexible and is suitable for both global and local modes of evaluations. Reference-independent evaluation is every well suited for local mode, but may work well for global as well, making reference alignments obsolete. This could be a very good thing provided difficulties to obtain correct structure alignments.

6.4.5 Why GDT_TS and LiveBench 3dscore Give Similar Results

During the experiments with GDT_TS and LiveBench 3dscore, we find out that the two scores give very similar results and conclusions (Figure 6.4 a & c). A further look at the GDT_TS and LiveBench 3dscore shows that the two scores actually give very similar values for the same pair of sequence alignment (Figure 6.15a) with a coefficient of determination (R^2) of 0.984. In order to understand this phenomenon, I take a closer look at their formulae. The formula of GDT_TS is

Equation 6.3 $$GDT\_TS = \frac{1}{L_{query}} \sum_{i=1}^{L_{ali}} GDT_i = \frac{1}{L_{query}} \sum_{1}^{L_{ali}} \left(\frac{n_1 + n_2 + n_4 + n_8}{4} \right)$$

which is a sum of a step function GDT_i (Figure 6.15b).
The formula of LiveBench 3dscore is

Equation 6.4 $$LiveBench\_3dscore = \frac{1}{L_{query}} \sum_{i=1}^{L_{ali}} LB3d_i = \frac{1}{L_{query}} \sum_{i=1}^{L_{ali}} \exp\left[-\ln 2 * \left(\frac{di}{3} \right)^2 \right]$$

which is a sum of a continuous function $LB3d_i$ (Figure 6.15b). The formula of function $LB3d_i$ is equivalent to formula $2^{-\left(\frac{d}{3}\right)^2}$, which is a continuous function. From Figure 6.15b we can see that $LB3d_i$ is a continuous simulation of the step function GDT_i. Therefore, there is no surprise that the two scores give very similar values.

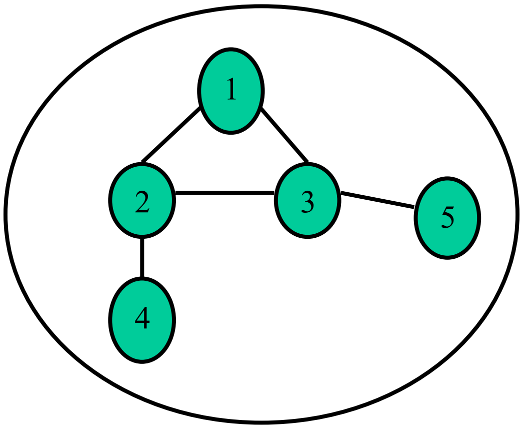
6.4.6 The Model Of Length Dependency Of Local GDT_TS Scores

The form of the model (power law) and the signs of the parameters for GDT_TS mean are consistent with the findings in the TMalign study (Zhang and Skolnick 2004), but our value of the parameters (Equation 6.2) are different from theirs ($mean(L) = 5.1L^{-0.74}$). This difference could be caused by different identity range of domain selections (ours: < 20%, theirs: < 30%), different superposition methods (ours: RMSD-optimal, theirs: TMalign), or different GDT_TS calculation modes (ours: local, theirs: global).

6.5 CONCLUSIONS

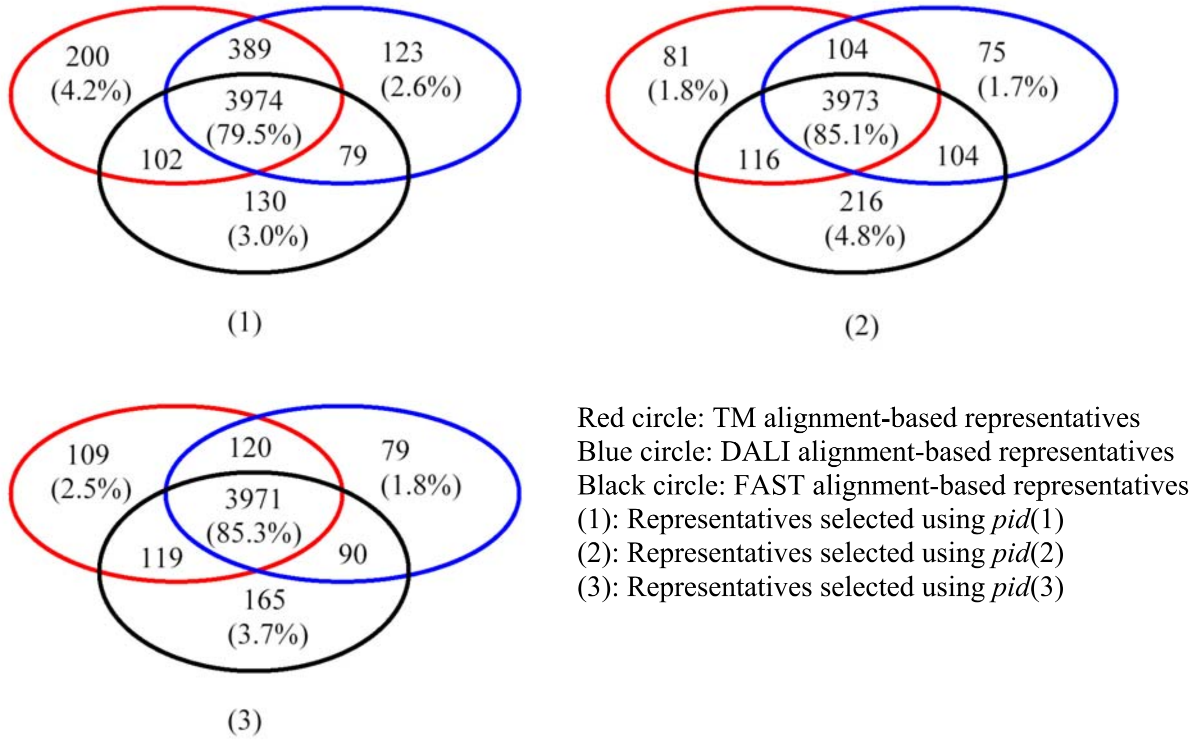
We have developed an automatic large-scale evaluation system aiming at systematically and comprehensively evaluating the structure modeling ability of sequence similarity search methods. We have first identified the pairwise identity calculation method that best reflects the evolutionary distance between protein domains and utilized this method to select 4147 representative SCOP protein domains as our testing set that have maximum 20% pairwise identities based on three types of structural alignments (DALI, TM, FAST). Both reference-dependent and reference-independent approaches are used to evaluate the fold recognition ability and alignment quality of different programs from global and local perspectives. For fold recognition ability (i.e. structure template quality) assessment, five structural and sequence similarity measures are found to be most effective and SVM technique is used to combine these measures. For alignment quality assessment, GDT_TS measure and the number of correct matches are utilized. Applying our evaluation system to six sequence similarity search programs show that our evaluation system is robust and helpful to shed light on the intrinsic properties of sequence similarity search programs.

Figure 6.1 Graphical Representation of Domain Relationship within A Superfamily



A graphical representation of relationships between domain sequences within a SCOP superfamily. Vertices 1-5 represent 5 domain sequences in this superfamily. An edge linking two vertices indicates the pairwise identity between these two domains is higher than cutoff. No edge between two vertices indicate the pairwise identity between the two is lower than cutoff.

Figure 6.2 Representatives Selected Using Different Structure Alignment Programs and Different Identity Measures



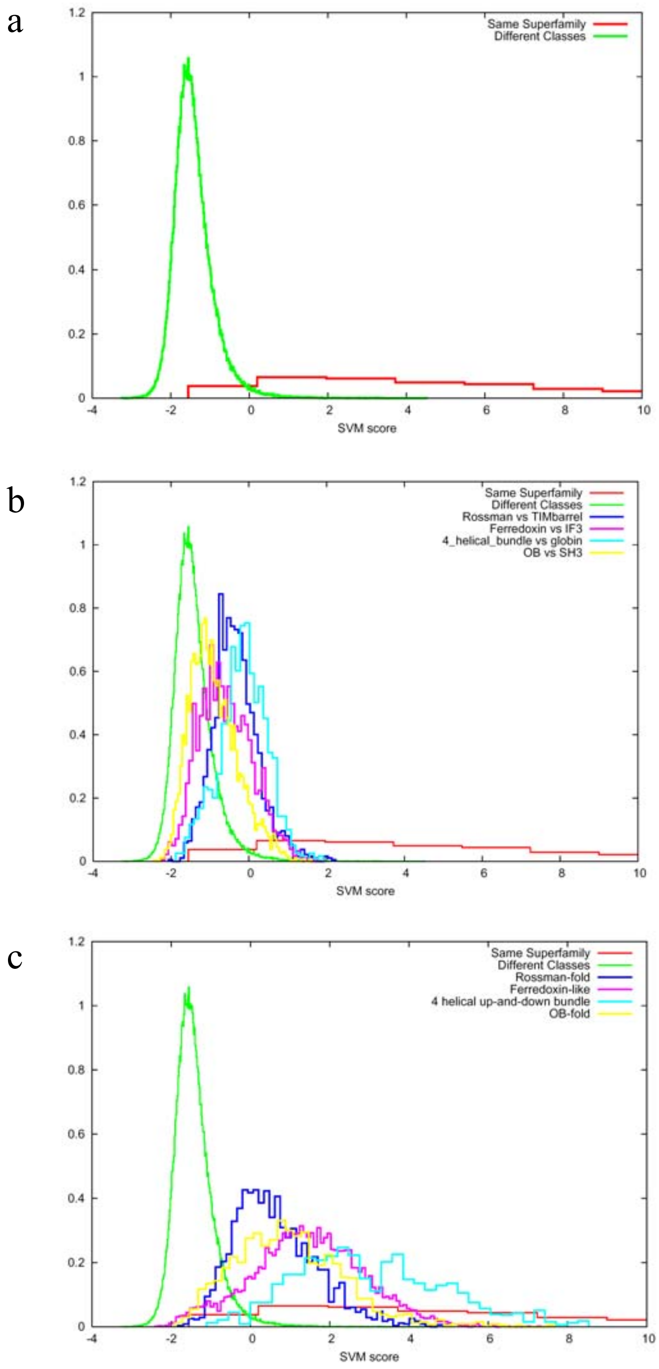
(1)

(2)

(3)

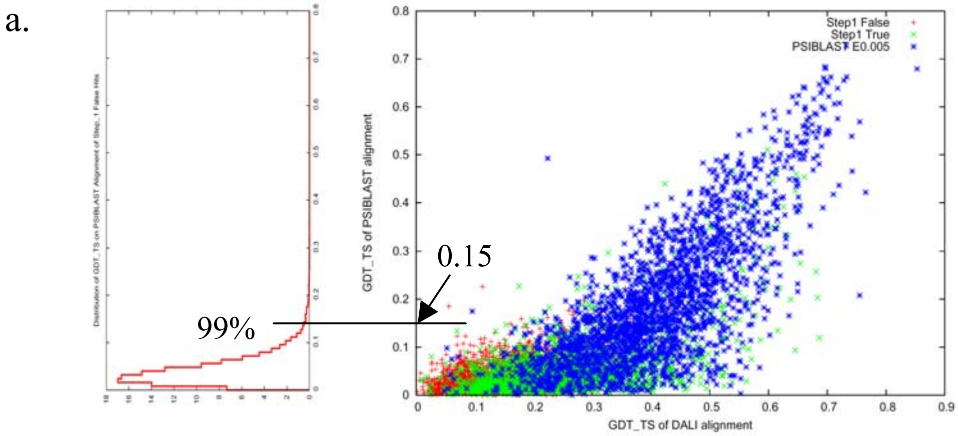
Red circle: TM alignment-based representatives
Blue circle: DALI alignment-based representatives
Black circle: FAST alignment-based representatives
(1): Representatives selected using *pid*(1)
(2): Representatives selected using *pid*(2)
(3): Representatives selected using *pid*(3)

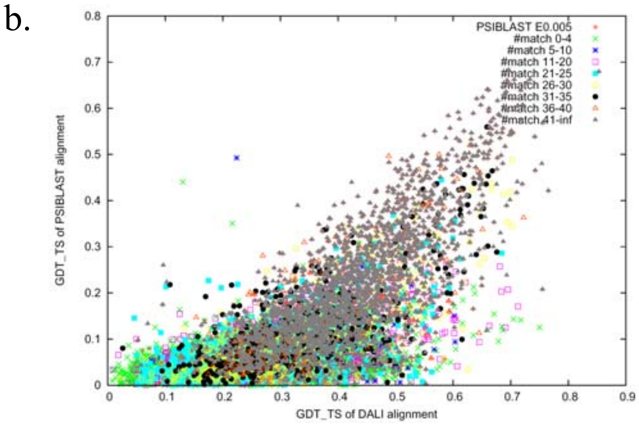
Figure 6.3 SVM Score Cutoff Selections for Overall Structure Template Quality

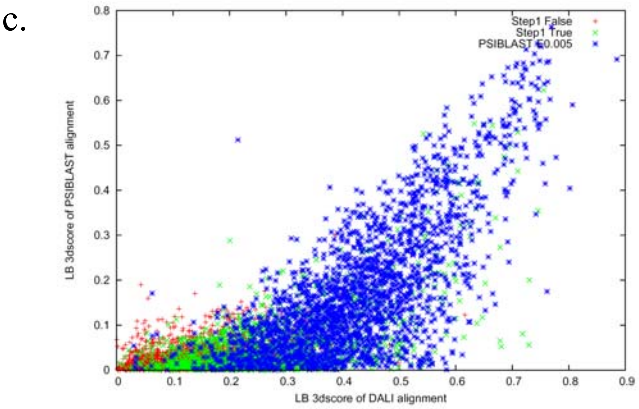


(a) SVM score distributions of the true (same superfamily) and false (different classes) hits based on SCOP classification. Best separation of the true and false is obtained by selected five features. (b) SVM score distributions of four inter-fold structure groups. The average (0.6) of the 95% percentiles of the four groups is taken as the high-cutoff. (c) SVM score distributions of four intra-fold structure groups. The average (-0.6) of the 5% percentiles of the four groups is taken as the low-cutoff.

Figure 6.4 Cutoff Selections for Alignment Quality

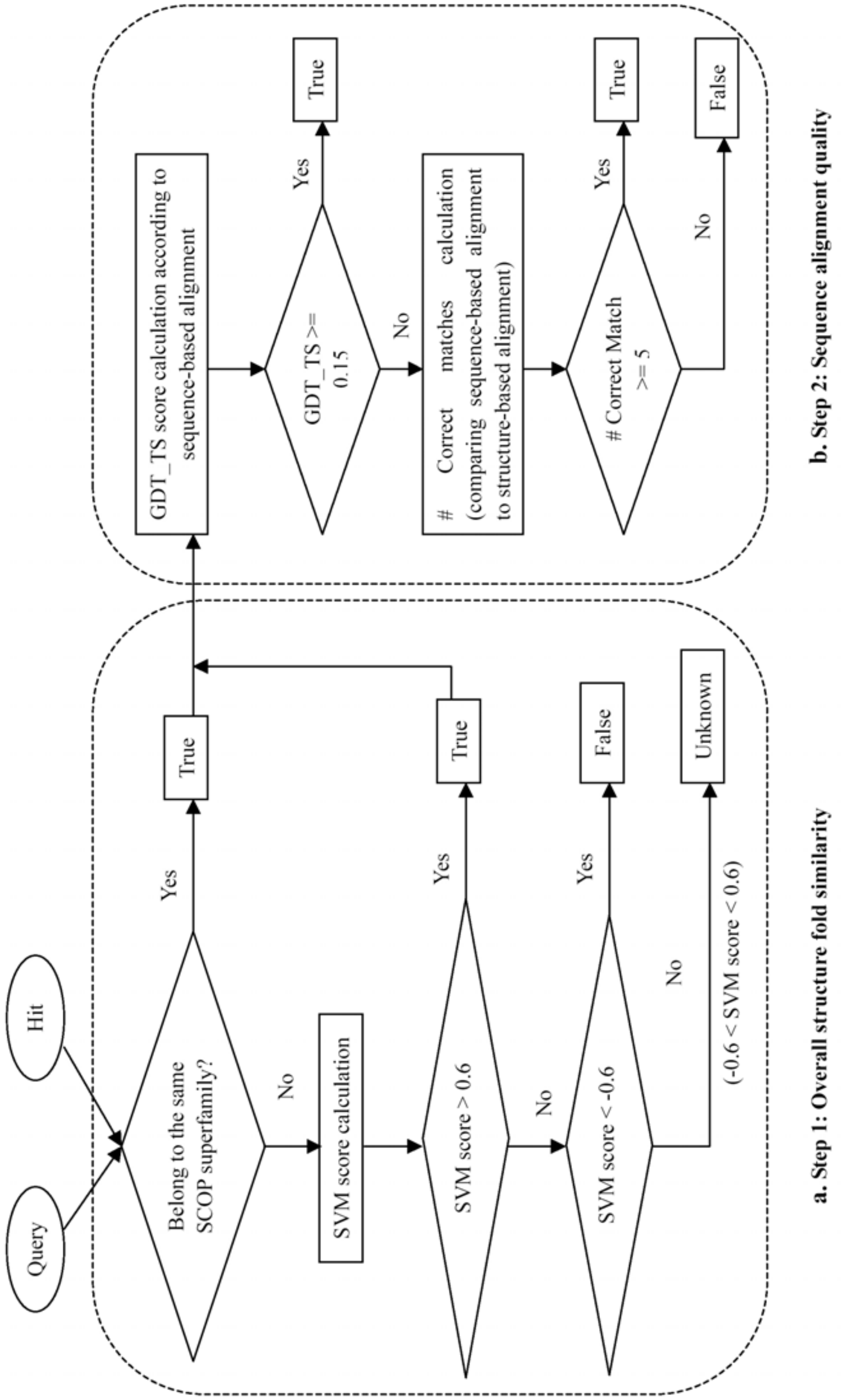






(a) GDT_TS scores of PSI-BLAST alignment vs. GDT_TS scores of structure-based DALI alignment of the step1 false domain pairs (red), step1 true domain pairs (green) and PSI-BLAST E-value less than 0.005 domain pairs (blue). (b) Coverage of the PSI-BLAST E-value less than 0.005 domain pairs by different ranges of number of correct matches. (c) LB 3dscores of PSI-BLAST alignments vs. LB 3dscors of DALI alignment. The same distributions as using GDT_TS scores as in (a).

Figure 6.5 Flowchart of Reference-dependent Evaluation System


Query
Hit
Belong to the same SCOP superfamily?
Yes
True
No
SVM score calculation
SVM score > 0.6
Yes
True
No
SVM score < -0.6
Yes
False
No
(-0.6 < SVM score < 0.6)
Unknown
GDT_TS score calculation according to sequence-based alignment
GDT_TS >= 0.15
Yes
True
No
# Correct matches calculation (comparing sequence-based alignment to structure-based alignment)
# Correct Match >= 5
Yes
True
No
False

a. Step 1: Overall structure fold similarity

b. Step 2: Sequence alignment quality

Figure 6.6 Length-dependency of GDT_TS

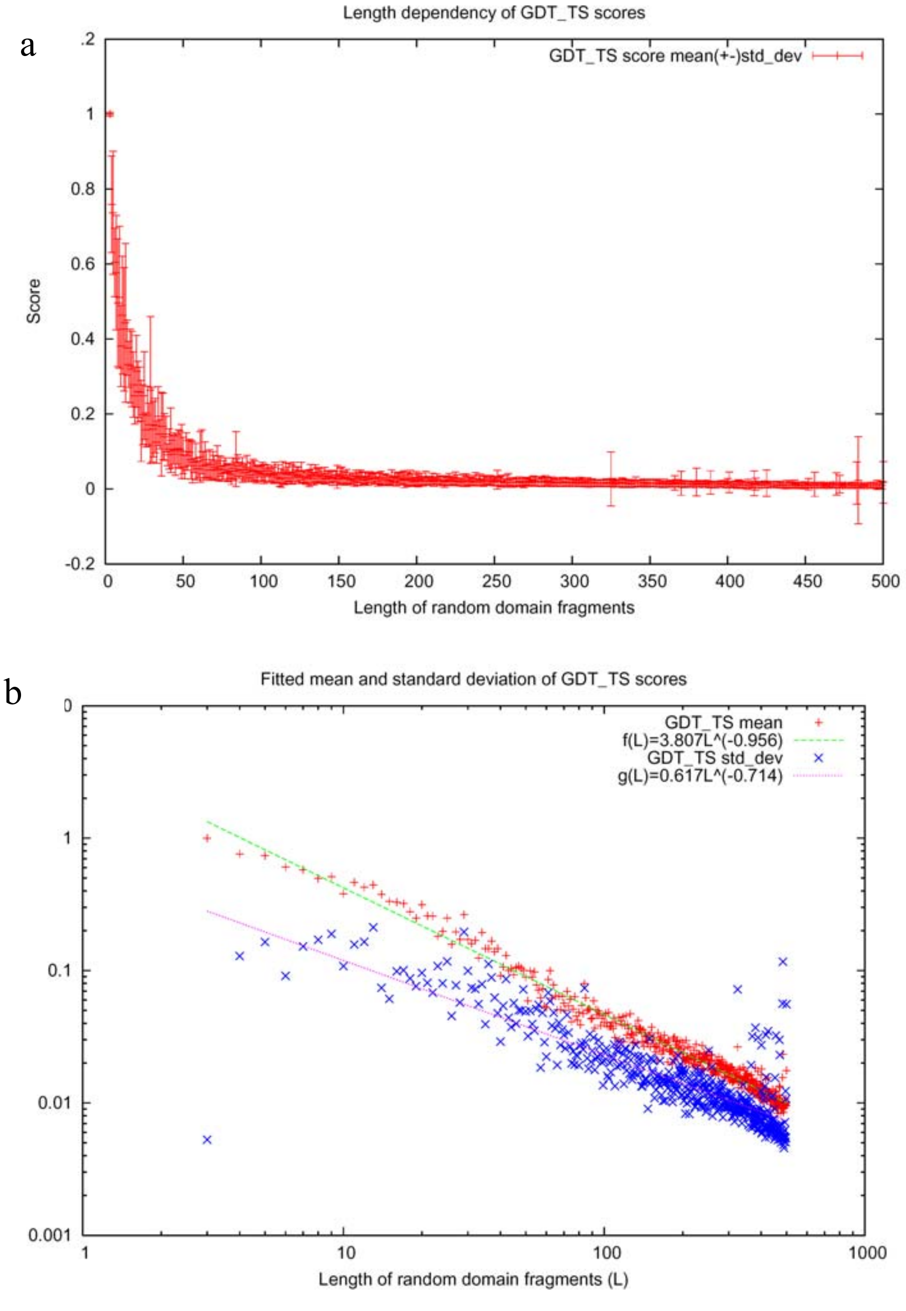
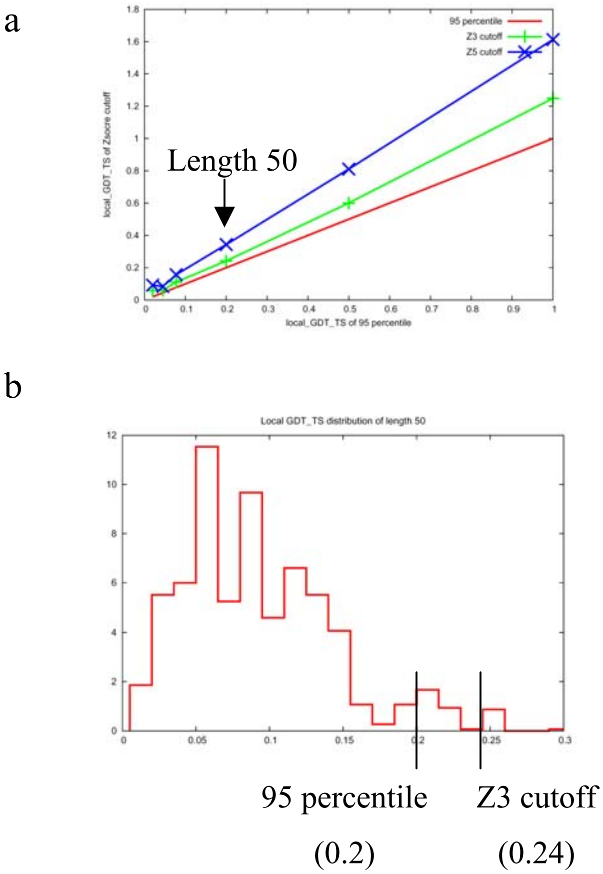


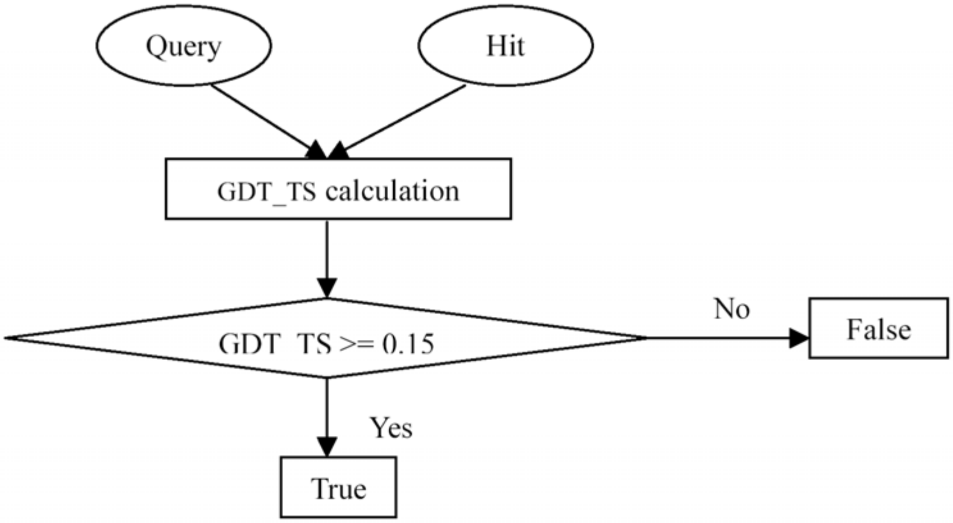
Figure 6.7 Local GDT_TS Cutoff Method Selection



(a) Comparison of the local GDT_TS values of Z-scores 3 (green) and 5 (blue) with that of 95 percentile of the local GDT_TS distribution for different local alignment lengths (5, 20, 50, 100, 200, 500 residues long). Both Z-score 3 and Z-score 5 cutoffs are more stringent than that of the 95 percentile of the distribution of local GDT_TS at a given length with Z-score 3 closer to the 95 percentiles. (b) Local GDT_TS distribution of length 50 (1000 samples). Z-score 3 cutoff value is more extreme than the 95 percentile value.

Figure 6.8 Flowchart of Reference-independent Evaluation System

a. Global mode



b. Local mode

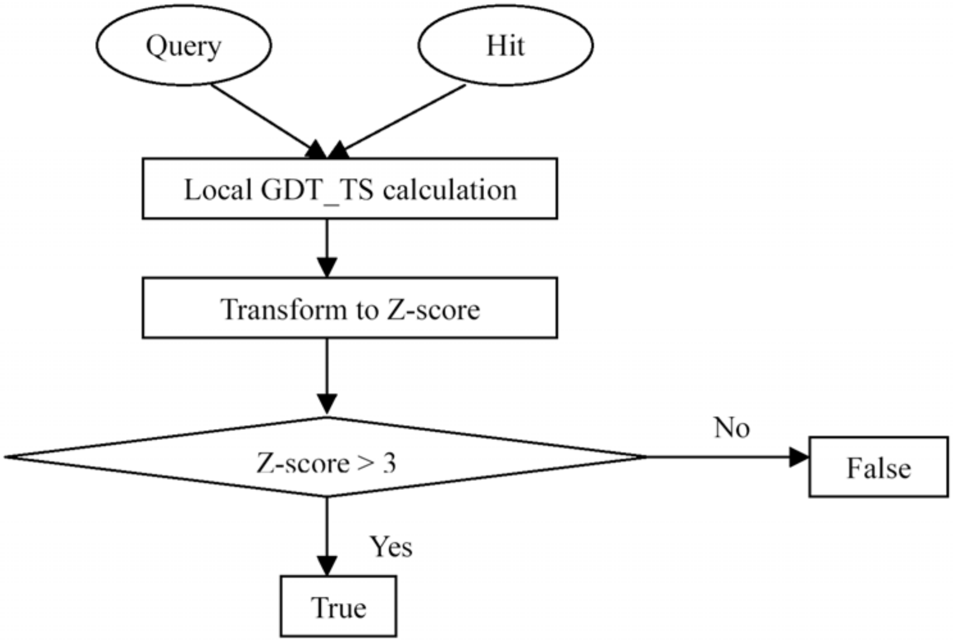
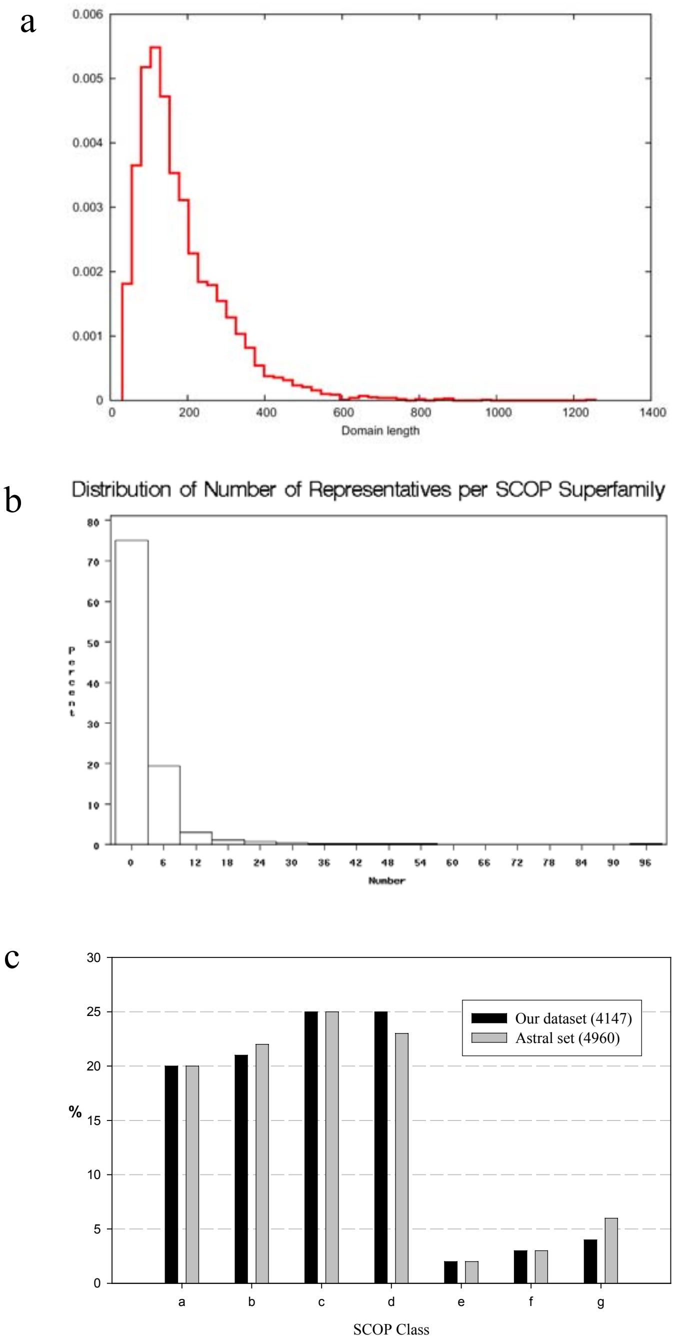
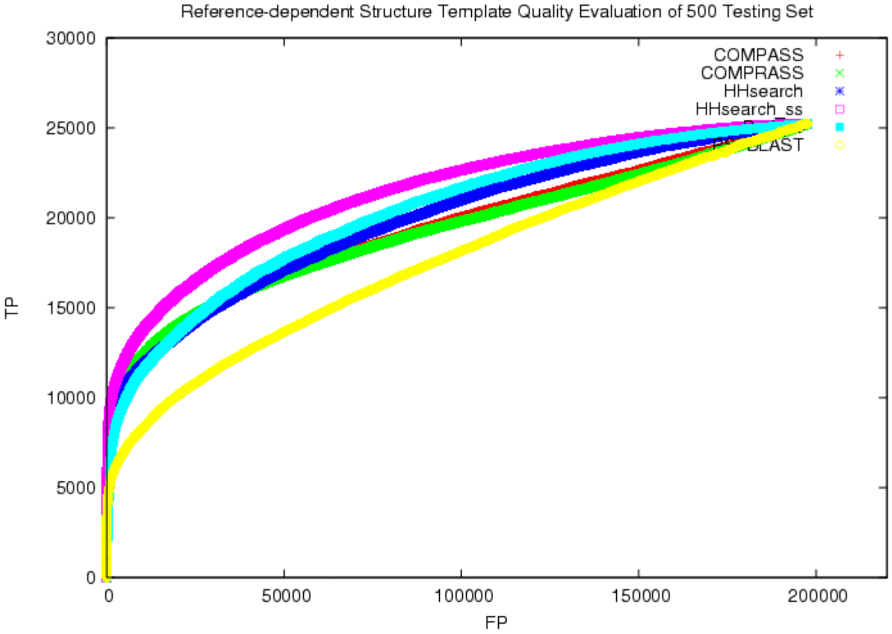


Figure 6.9 Distributions of Representative Domains



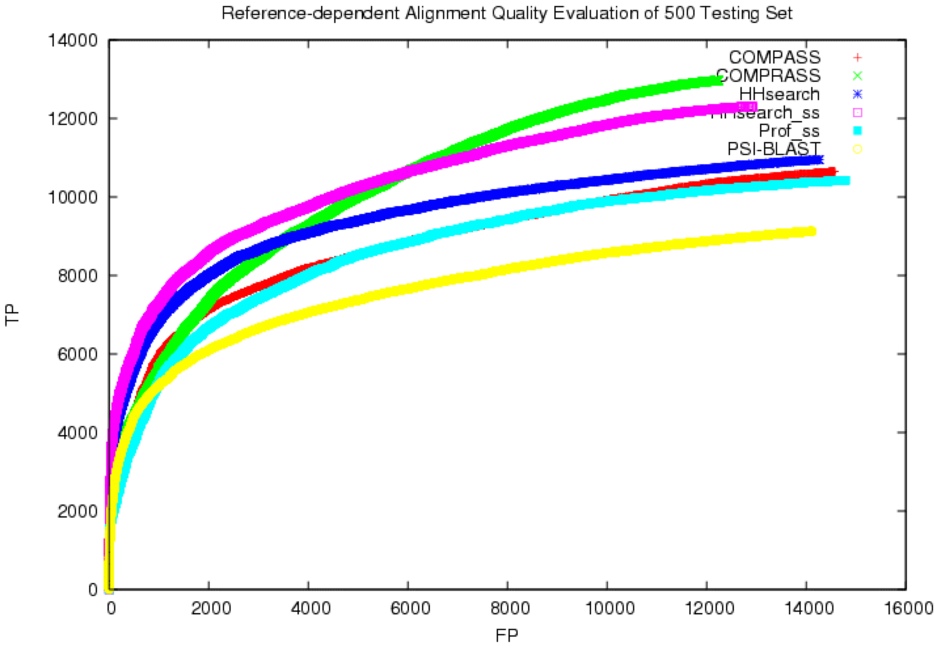
(a) Distribution of domain lengths of the representative dataset. (b) Distribution of number of representatives per SCOP superfamily of the representative dataset. (c) Distribution of number of representatives per SCOP Class. The black bar shows the distribution of our representative dataset that contains 4147 domain sequences. The grey bar shows the distribution of Astral SCOP20 set which contains 4960 domain sequences. The x-axis shows the abbreviate names of the SCOP Classes. a: all alpha proteins; b: all beta proteins; c: alpha/beta proteins; d: alpha+beta proteins; e: multi-domain proteins; f: membrane and cell surface proteins and peptides; g: small proteins.

Figure 6.10 Reference-dependent Structure Template Quality Evaluation of Various Sequence Alignment Programs on the Testing Set



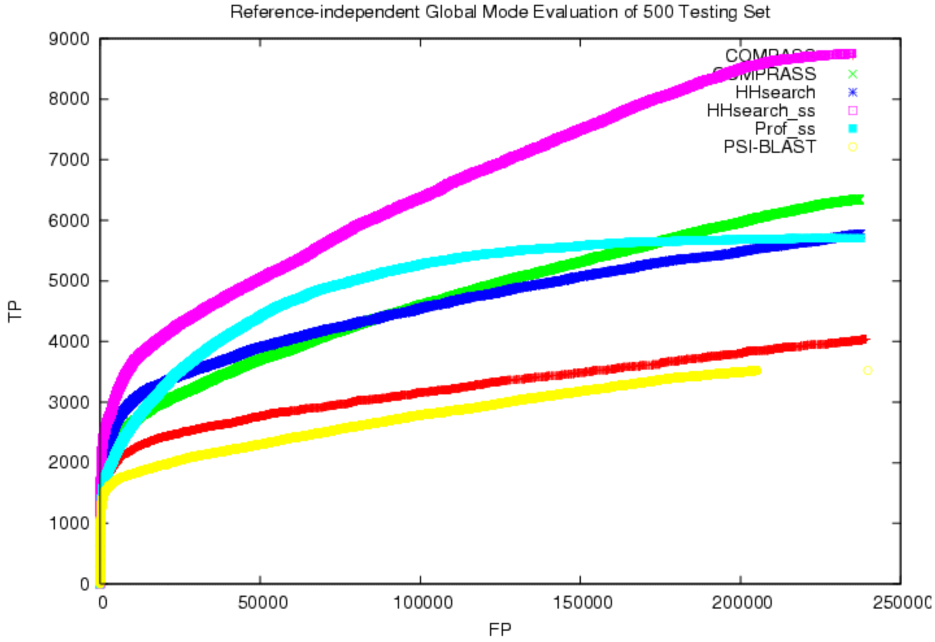
Different colors of the ROC curves indicate different programs.

Figure 6.11 Reference-dependent Alignment Quality Evaluation of Various Sequence Alignment Programs on the Testing Set



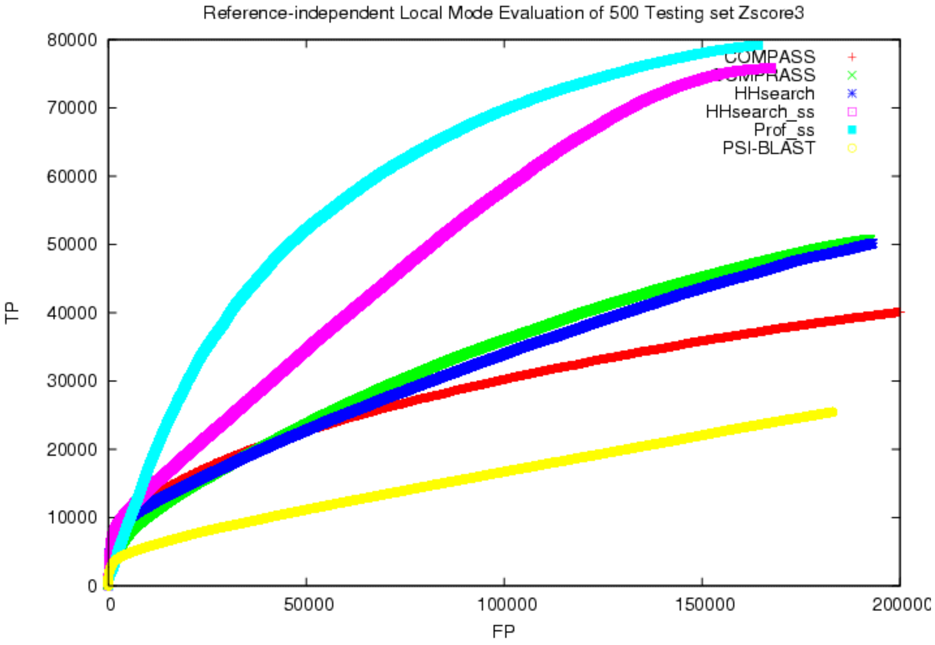
Different colors of the ROC curves indicate different programs.

Figure 6.12 Reference-independent Global Mode Evaluation of Various Sequence Alignment Programs on the Testing Set



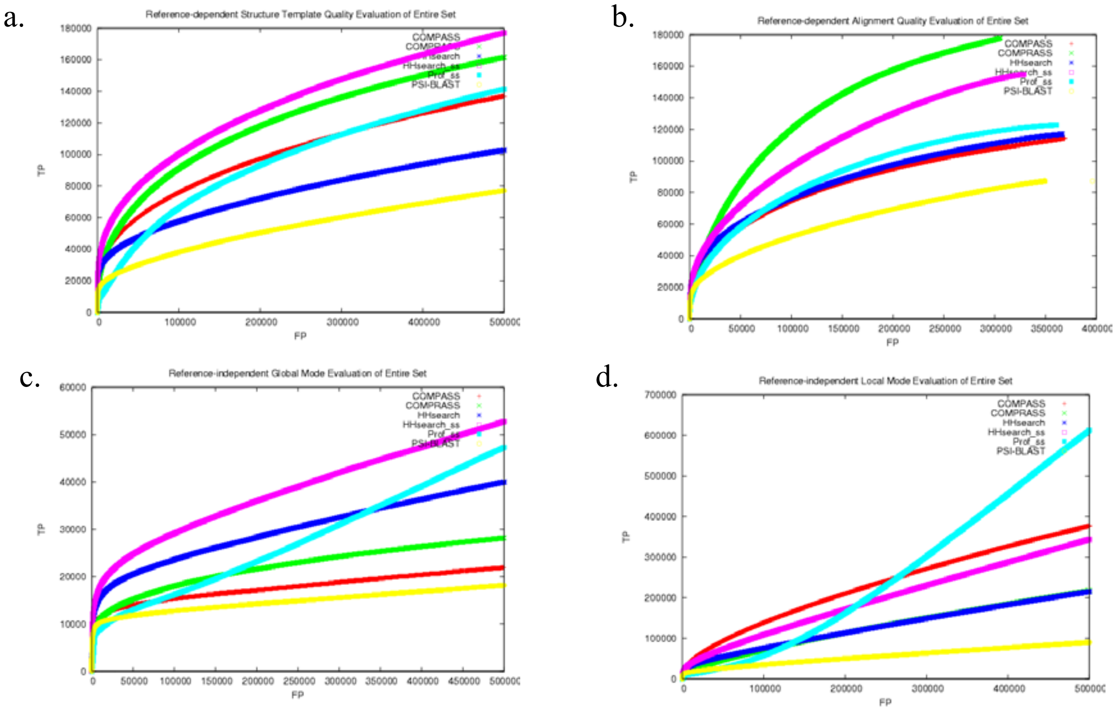
Different colors of the ROC curves indicate different programs.

Figure 6.13 Reference-independent Local Mode Evaluation of Various Sequence Alignment Programs on the Testing Set



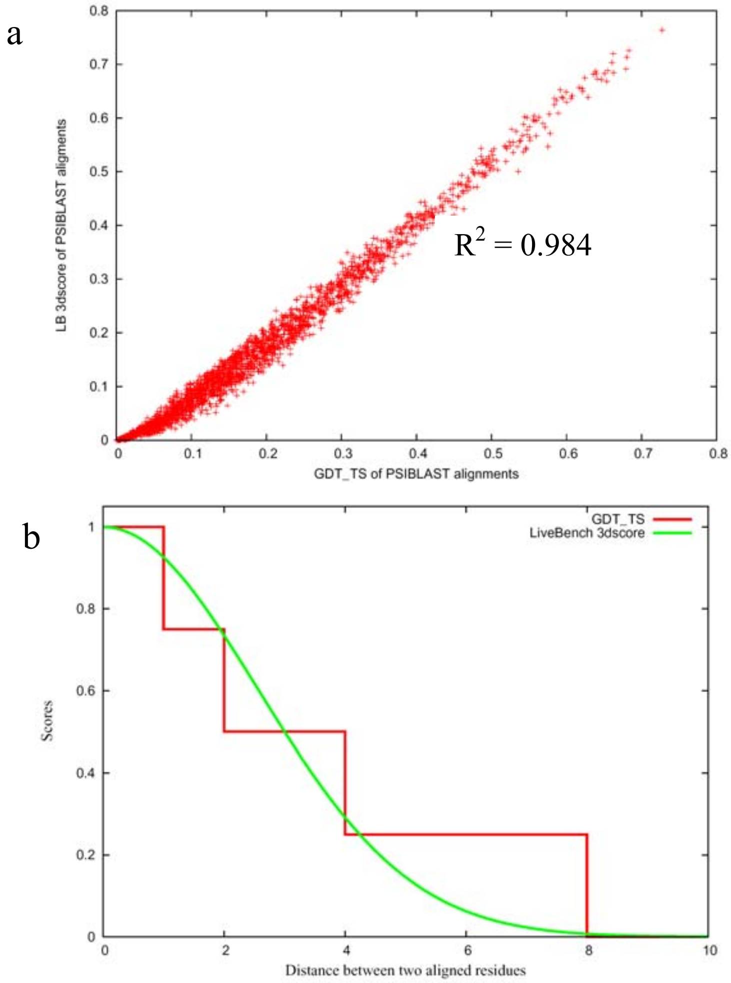
Different colors of the ROC curves indicate different programs.

Figure 6.14 Evaluation Results of Various Sequence Alignment Programs on the Entire Representative Dataset



(a) Reference-dependent structure template evaluation results. (b) Reference-dependent alignment quality evaluation results. (c) Reference-independent global mode evaluation results. (d) Reference-independent local mode evaluation results. Different colors of the ROC curves indicate different programs.

Figure 6.15 Comparison of GDT_TS and LiveBench 3dscore Functions



(a): GDT_TS and LiveBench 3d score give similar score values. (b). Function of GDT_i (red) and $LB3d_i$ (green). The lines indicate the changes of function values with the change of distances between two aligned residues.

$$GDT_i = \frac{n_1 + n_2 + n_4 + n_8}{4}$$

$$LB3d_i = 2^{-\left(\frac{d}{3}\right)^2}$$

Table 6.1 SVM Score Cutoffs

a. High-cutoff

Structural class	Representative problematic fold pair	SVM score at 95% percentile
α/β	Rossmann-fold vs. TIM barrel	0.7
α+β	Ferredoxin-like vs. IF3-like	0.6
All α	Four-helical up-and-down bundle vs. globin-like	0.8
All β	OB-fold vs. SH3-like barrel	0.3
Avg		**0.6**

b. Low-cutoff

Structural class	Representative fold	SVM score at 5% percentile
α/β	Rossmann-fold	-0.8
α+β	Ferredoxin-like	-1.0
All α	Four-helical up-and-down bundle	0.4
All β	OB-fold	-1.1
Avg		**-0.6**

CHAPTER 7:
Summary and Future Directions

In an attempt to explore and utilize the sequence-structure-function relationships of proteins, this dissertation work mainly focused on algorithmic development to address homology detection and utilization related issues, including more powerful homology detection methods, structure modeling ability evaluations, and positional correlation based functional predictions. The developed algorithms and methods are generally applicable to all protein families. Case studies of structure prediction and structure classification are also carried out to address problems in specific protein family or groups of families.

7.1 CONCLUDING REMARKS: STRUCTURE PREDICTION OF GYRASE A C-TERMINAL DOMAIN

7.1.1 Project Summary

A structure prediction of the C-terminal domain of Gyrase A (GyrA) and topoisomerase IV (ParC) is presented in Chapter 2. The C-terminal domain of GyrA/ParC was the largest piece of topoisomerase sequence without available structural information at the time the prediction was made. Using extensive sequence and structure analysis of the GyrA/ParC C-terminal domain and regulator of chromosome condensation (RCC1), including sequence similarity search, multiple sequence alignment, hydrophobicity analysis, secondary structure prediction, and fold recognition, we infer homology between these proteins and therefore predict the structural fold and functional implications of the GyrA/ParC C-terminal domain. The fold prediction is later verified by experimental data.

7.1.2 Applications and Utility

This chapter illustrates a case study of homology-based structure prediction. The results of the structure prediction and functional implications are directly beneficial to researchers working with DNA topoisomerases. Since most of the current methods and techniques for protein structure prediction are used in this project, the process of this project could serve as a template for researchers who wants to perform structure prediction/modeling for their own proteins.

7.2 CONCLUDING REMARKS: STRUCTURE CLASSIFICATION OF THIOREDOXIN-LIKE FOLD PROTEINS

7.2.1 Project Summary

A hierarchical structure classification of thioredoxin-like fold proteins is presented in Chapter 3. The thioredoxin-like fold is defined and protein domains containing the thioredoxin-like fold are identified through extensive structural search and are classified into fold groups and evolutionary families through sequence, structure and functional analysis. The characteristic structural or functional features of each evolutionary family are described in detail. A multiple structural alignment on ninety representatives is performed. Analysis of active site locations is carried out.

7.2.2 Applications and Utility

This structure classification has multiple benefits. The thioredoxin-like fold is defined firstly based on structural consensus of thioredoxin homologs and explicit usage of circular permutations, and therefore is useful to clarify fold definitions. The resulting definition is more inclusive compared to exiting classifications, which helps identify previously unrecognized similarities between proteins that are newly brought together. Circular permutation analysis also helps reveal potential functional/packing unit. Furthermore, because the nature of this structure classification emphasizes on convergent evolution of structural folds, a thorough study of these protein domains may aid in understanding of the physico-chemical principles behind protein structures, which in turn could help to address problems such as protein folding and structure-functional predictions. The structure-based multiple sequence alignment of the thioredoxin-like fold proteins offers information on protein sequence-structure relationships and can be employed in protein structure predictions. Analysis of active site locations offers useful information on protein structure-function relationships and can be employed in protein functional predictions.

7.3 CONCLUDING REMARKS: POSITIONAL CORRELATION ANALYSIS ALGORITHM

7.3.1 Project Summary

The development of a software package, PCOAT (Positional Correlation Analysis Tool), is presented in Chapter 4. PCOAT has been developed to perform positional correlation analysis for protein multiple sequence alignments. Different statistical methods have been implemented to detect highly correlated position pairs, amino acid pairs, individual positions, and networks of correlated positions. Multiple sequence weighting and sampling methods have been developed to eliminate background correlations caused by phylogeny and stochastic events.

7.3.2 Applications and Utility

Because correlations between protein positions often arise for structural or functional reasons, such as stabilizing local contact or affecting protein functions through networks of interactions, PCOAT should be useful and convenient for researchers to predict positions or residues of structurally or functionally important interactions in their protein families.

PCOAT runs relatively fast and is suitable for analyzing alignments containing large number of sequences.

7.4 CONCLUDING REMARKS: SEQUENCE SIMILARITY SEARCH METHOD

7.4.1 Project Summary

The development of a more sensitive sequence similarity search method is presented in Chapter 5. With increased structure modeling and homology detection abilities as the goals, this method makes use of the predicted secondary structure information and combines it with sequence profiles. Substitution matrices of predicted secondary structure elements and amino acids are calculated and used in the scoring system developed for measuring sequence-secondary structure similarities. The parameters of a statistical model are fitted in order to estimate the statistical significance of resulting scores.

7.4.2 Applications and Utility

This method can be of use to both computational biologist and experimental researchers. With increased sensitivity for homology detection ability, this method can find more remotely similar homologs that can serve as structure template for newly discovered or poorly studies proteins that are distant from other proteins in structural space. In addition, because homologous proteins usually preserve the same general biochemical function, making a rough functional prediction is possible using this method for newly discovered proteins. Further more, with increased alignment quality, this method provides longer and better alignments that are suitable for structure modeling purposes. Applying to the PFAM families shows that this method can be used to identify previous unrecognized similarities between protein families, as well as to directly identify homologs that were previously identified only through transitive method.

7.5 CONCLUDING REMARKS: EVALUATION SYSTEM FOR SEQUENCE SIMILARITY SEARCH METHODS

7.5.1 Project Summary

A comprehensive evaluation system for the structure modeling abilities of sequence similarity search methods is presented in Chapter 6. A large, non-biased representative protein domain set is first selected to serve as the testing set. Different sequence and structure similarity measures are then combined to assess the sensitivity and specificity of different programs. The evaluation procedure makes automatic assessments for both fold recognition abilities and alignment qualities from global and local perspectives using both reference-dependent and reference-independent approaches.

7.5.2 Applications and Utility

This evaluation system is of particular interest to the protein structure modeling community, both to the method developers and the users. It serves as an instrument to benchmark the structure modeling abilities of different sequence similarity search and alignment methods. Researchers developing sequence similarity search methods need to test and compare their performances all the time. Such an evaluation system helps the developers to understand the achievements and limitations of their programs and of the field, as well as helps the users to choose the appropriate programs to use for their purposes.

APPENDIX A
FORMULAE

A.1 Other variations of pairwise percentage identity calculation:

1) Combined id2: $pid(4) = \dfrac{N_{id} + N_{id\_unali}}{L_{ali} + L_{sum\_shoter\_unali}}$

2) Variation of $pid(3)$ and $pid(4)$: $pid(5) = \dfrac{N_{id} + L_{sum\_shorter\_unali} * pid_{random}}{L_{ali} + L_{sum\_shorter\_unali}}$

3) Variation of $pid(2)$: $pid(6) = \dfrac{N_{id}}{L_{ali} + L_{sum\_shorter\_unali}}$

BIBLIOGRAPHY

Afonnikov, D. A., D. Y. Oshchepkov, et al. (2001). "Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions." Bioinformatics **17**(11): 1035-46.

Altschul, S. F. (1991). "Amino acid substitution matrices from an information theoretic perspective." J Mol Biol **219**(3): 555-65.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Anantharaman, V. and L. Aravind (2006). "Diversification of catalytic activities and ligand interactions in the protein fold shared by the sugar isomerases, eIF2B, DeoR transcription factors, acyl-CoA transferases and methenyltetrahydrofolate synthetase." J Mol Biol **356**(3): 823-42.

Aravind, L., D. D. Leipe, et al. (1998). "Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins." Nucleic Acids Res **26**(18): 4205-4213.

Arner, E. S. and A. Holmgren (2000). "Physiological functions of thioredoxin and thioredoxin reductase." Eur J Biochem **267**(20): 6102-9.

Atchley, W. R., K. R. Wollenberg, et al. (2000). "Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis." Mol Biol Evol **17**(1): 164-78.

Attwood, T. K., M. E. Beck, et al. (1994). "PRINTS--a database of protein motif fingerprints." Nucleic Acids Res **22**(17): 3590-6.

Attwood, T. K., P. Bradley, et al. (2003). "PRINTS and its automatic supplement, prePRINTS." Nucleic Acids Res **31**(1): 400-2.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32**(Database issue): D138-41.

Berger, J. M., D. Fass, et al. (1998). "Structural similarities between topoisomerases that cleave one or both DNA strands." Proc Natl Acad Sci U S A **95**(14): 7876-7881.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-42.

Brenner, S. E., P. Koehl, et al. (2000). "The ASTRAL compendium for protein structure and sequence analysis." Nucleic Acids Res **28**(1): 254-6.

Bryson, K., L. J. McGuffin, et al. (2005). "Protein structure prediction servers at University College London." Nucleic Acids Res **33**(Web Server issue): W36-8.

Bujnicki, J. M. (2002). "Sequence permutations in the molecular evolution of DNA methyltransferases." BMC Evol Biol **2**(1): 3.

Burley, S. K. (2000). "An overview of structural genomics." Nat Struct Biol **7 Suppl**: 932-4.

Burley, S. K., S. C. Almo, et al. (1999). "Structural genomics: beyond the human genome project." Nat Genet **23**(2): 151-7.

Caron, P. R. and J. C. Wang (1994). "Appendix. II: Alignment of primary sequences of DNA topoisomerases." Adv Pharmacol: 271-297.

Chandonia, J. M., G. Hon, et al. (2004). "The ASTRAL Compendium in 2004." Nucleic Acids Res **32**(Database issue): D189-92.

Chandonia, J. M. and M. Karplus (1999). "New methods for accurate prediction of protein secondary structure." Proteins **35**(3): 293-306.

Chandonia, J. M., N. S. Walker, et al. (2002). "ASTRAL compendium enhancements." Nucleic Acids Res **30**(1): 260-3.

Cheek, S., Y. Qi, et al. (2004). "4SCOPmap: automated assignment of protein structures to evolutionary superfamilies." BMC Bioinformatics **5**: 197.

Chook, Y. M., J. V. Gray, et al. (1994). "The monofunctional chorismate mutase from Bacillus subtilis. Structure determination of chorismate mutase and its complexes with a transition state analog and prephenate, and implications for the mechanism of the enzymatic reaction." J Mol Biol **240**(5): 476-500.

Chook, Y. M., H. Ke, et al. (1993). "Crystal structures of the monofunctional chorismate mutase from Bacillus subtilis and its complex with a transition state analog." Proc Natl Acad Sci U S A **90**(18): 8600-3.

Chou, P. Y. and G. D. Fasman (1978). "Empirical predictions of protein conformation." Annu Rev Biochem **47**: 251-76.

Chung, R. and G. Yona (2004). "Protein family comparison using statistical models and predicted structural information." BMC Bioinformatics **5**: 183.

Corbett, K. D., R. K. Shultzaberger, et al. (2004). "The C-terminal domain of DNA gyrase A adopts a DNA-bending beta-pinwheel fold." Proc Natl Acad Sci U S A **101**(19): 7293-8.

Crowder, S., J. Holton, et al. (2001). "Covariance analysis of RNA recognition motifs identifies functionally linked amino acids." J Mol Biol **310**(4): 793-800.

Cuff, J. A. and G. J. Barton (2000). "Application of multiple sequence alignment profiles to improve protein secondary structure prediction." Proteins **40**(3): 502-511.

Das, K. C. (2004). "Thioredoxin system in premature and newborn biology." Antioxid Redox Signal **6**(1): 177-84.

Dayhoff, M. O. (1978). Atlas of protein sequence and structure. Silver Spring, Md.,, National Biomedical Research Foundation.

Deibler, R. W., S. Rahmati, et al. (2001). "Topoisomerase IV, alone, unknots DNA in E. coli." Genes Dev **15**(6): 748-761.

Dietmann, S. and L. Holm (2001). "Identification of homology in protein structure classification." Nat Struct Biol **8**(11): 953-7.

Ermler, U., M. Merckel, et al. (1997). "Formylmethanofuran: tetrahydromethanopterin formyltransferase from Methanopyrus kandleri - new insights into salt-dependence and thermostability." Structure **5**(5): 635-46.

Eyrich, V. A., M. A. Marti-Renom, et al. (2001). "EVA: continuous automatic evaluation of protein structure prediction servers." Bioinformatics **17**(12): 1242-3.

Fischer, D. (2000). Hybrid Fold Recognition: combining sequence derived properties with evolutionary information. Pacific Symp. Biocomputing, Hawaii.

Fischer, D., L. Rychlewski, et al. (2003). "CAFASP3: the third critical assessment of fully automated structure prediction methods." Proteins **53 Suppl 6**: 503-16.

Fitch, W. M. (2000). "Homology a personal view on some of the problems." Trends Genet **16**(5): 227-31.

Ginalski, K., N. V. Grishin, et al. (2005). "Practical lessons from protein structure prediction." Nucleic Acids Res **33**(6): 1874-91.

Ginalski, K., J. Pas, et al. (2003). "ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure." Nucleic Acids Res **31**(13): 3804-7.

Ginalski, K., M. von Grotthuss, et al. (2004). "Detecting distant homology with Meta-BASIC." Nucleic Acids Res **32**(Web Server issue): W576-81.

Gitschier, J., B. Moffat, et al. (1998). "Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATPase." Nat Struct Biol **5**(1): 47-54.

Gong, W., M. O'Gara, et al. (1997). "Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment." Nucleic Acids Res **25**(14): 2702-15.

Greasley, S. E., P. Horton, et al. (2001). "Crystal structure of a bifunctional transformylase and cyclohydrolase enzyme in purine biosynthesis." Nat Struct Biol **8**(5): 402-6.

Grishin, N. V. (2001). "KH domain: one motif, two folds." Nucleic Acids Res **29**(3): 638-43.

Hadley, C. and D. T. Jones (1999). "A systematic comparison of protein structure classifications: SCOP, CATH and FSSP." Structure Fold Des **7**(9): 1099-112.

Helgstrand, C., W. R. Wikoff, et al. (2003). "The refined structure of a protein catenane: the HK97 bacteriophage capsid at 3.44 A resolution." J Mol Biol **334**(5): 885-99.

Henikoff, S. and J. G. Henikoff (1991). "Automated assembly of protein blocks for database searching." Nucleic Acids Res **19**(23): 6565-72.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-9.

Henikoff, S. and J. G. Henikoff (1994). "Position-based sequence weights." J Mol Biol **243**(4): 574-8.

Henikoff, S., J. G. Henikoff, et al. (1999). "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations." Bioinformatics **15**(6): 471-9.

Hennecke, J., P. Sebbel, et al. (1999). "Random circular permutation of DsbA reveals segments that are essential for protein folding and stability." J Mol Biol **286**(4): 1197-215.

Hiasa, H., D. O. Yousef, et al. (1996). "DNA strand cleavage is required for replication fork arrest by a frozen topoisomerase-quinolone-DNA ternary complex." J Biol Chem **271**(42): 26424-26429.

Holm, L. and J. Park (2000). "DaliLite workbench for protein structure comparison." Bioinformatics **16**(6): 566-7.

Holm, L. and C. Sander (1995). "Dali: a network tool for protein structure comparison." Trends Biochem Sci **20**(11): 478-80.

Holm, L. and C. Sander (1996). "The FSSP database: fold classification based on structure-structure alignment of proteins." Nucleic Acids Res **24**(1): 206-9.

Holm, L. and C. Sander (1996). "Mapping the protein universe." Science **273**(5275): 595-603.

Holm, L. and C. Sander (1998). "Dictionary of recurrent domains in protein structures." Proteins **33**(1): 88-96.

Holmgren, A. (1995). "Thioredoxin structure and mechanism: conformational changes on oxidation of the active-site sulfhydryls to a disulfide." Structure **3**(3): 239-43.

Hsieh, T. J., L. Farh, et al. (2004). "Structure of the topoisomerase IV C-terminal domain: a broken beta-propeller implies a role as geometry facilitator in catalysis." J Biol Chem **279**(53): 55587-93.

Huber, T., A. J. Russell, et al. (1999). "Sausage: protein threading with flexible force fields." Bioinformatics **15**(12): 1064-5.

Hulo, N., A. Bairoch, et al. (2006). "The PROSITE database." Nucleic Acids Res **34**(Database issue): D227-30.

Interthal, H., J. J. Pouliot, et al. (2001). "The tyrosyl-DNA phosphodiesterase Tdp1 is a member of the phospholipase D superfamily." Proc Natl Acad Sci U S A **98**(21): 12009-14.

Iuchi, S. (2001). "Three classes of C2H2 zinc finger proteins." Cell Mol Life Sci **58**(4): 625-35.

Jeltsch, A. (1999). "Circular permutations in the molecular evolution of DNA methyltransferases." J Mol Evol **49**(1): 161-4.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. Advances in kernel methods : support vector learning. B. Schölkopf, C. J. C. Burges and A. J. Smola. Cambridge, Mass., MIT Press.

Johansson, E., N. Mejlhede, et al. (2002). "Crystal structure of the tetrameric cytidine deaminase from Bacillus subtilis at 2.0 A resolution." Biochemistry **41**(8): 2563-70.

Jones, D. T. (1999). "GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences." J Mol Biol **287**(4): 797-815.

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." J Mol Biol **292**(2): 195-202.

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-637.

Karlin, S. and S. F. Altschul (1990). "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." Proc Natl Acad Sci U S A **87**(6): 2264-8.

Karplus, K., C. Barrett, et al. (1998). "Hidden Markov models for detecting remote protein homologies." Bioinformatics **14**(10): 846-56.

Kelley, L. A., R. M. MacCallum, et al. (2000). "Enhanced genome annotation using structural profiles in the program 3D-PSSM." J Mol Biol **299**(2): 499-520.

Kemp, L. E., C. S. Bond, et al. (2002). "Structure of 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase: an essential enzyme for isoprenoid biosynthesis and target for antimicrobial drug development." Proc Natl Acad Sci U S A **99**(10): 6591-6.

Kinch, L. N., J. O. Wrabl, et al. (2003). "CASP5 assessment of fold recognition target predictions." Proteins **53 Suppl 6**: 395-409.

Kishida, H., T. Wada, et al. (2003). "Structure and catalytic mechanism of 2-C-methyl-D-erythritol 2,4-cyclodiphosphate (MECDP) synthase, an enzyme in the non-mevalonate pathway of isoprenoid synthesis." Acta Crystallogr D Biol Crystallogr **59**(Pt 1): 23-31.

Knight, S. W. and D. S. Samuels (1999). "Natural synthesis of a DNA-binding protein from the C-terminal domain of DNA gyrase A in Borrelia burgdorferi." Embo J **18**(17): 4875-4881.

Kobe, B., I. G. Jennings, et al. (1999). "Structural basis of autoregulation of phenylalanine hydroxylase." Nat Struct Biol **6**(5): 442-8.

Kolodny, R., P. Koehl, et al. (2005). "Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures." J Mol Biol **346**(4): 1173-88.

Kontou, M., R. D. Will, et al. (2004). "Thioredoxin, a regulator of gene expression." Oncogene.

Kraulis, P. J. (1991). "MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures." J. Appl. Crystallogr. **24**: 946-950.

Larson, S. M., A. A. Di Nardo, et al. (2000). "Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions." J Mol Biol **303**(3): 433-46.

Lattman, E. E. (2005). "Sixth meeting on the critical assessment of techniques for protein structure prediction." Proteins **61**(S7): 1-2.

Letunic, I., R. R. Copley, et al. (2006). "SMART 5: domains in the context of genomes and networks." Nucleic Acids Res **34**(Database issue): D257-60.

Li, C., T. J. Kappock, et al. (1999). "X-ray crystal structure of aminoimidazole ribonucleotide synthetase (PurM), from the Escherichia coli purine biosynthetic pathway at 2.5 A resolution." Structure Fold Des **7**(9): 1155-66.

Lo Conte, L., B. Ailey, et al. (2000). "SCOP: a structural classification of proteins database." Nucleic Acids Res **28**(1): 257-9.

Lo Conte, L., S. E. Brenner, et al. (2002). "SCOP database in 2002: refinements accommodate structural genomics." Nucleic Acids Res **30**(1): 264-7.

Lowe, J., H. Li, et al. (2001). "Refined structure of alpha beta-tubulin at 3.5 A resolution." J Mol Biol **313**(5): 1045-57.

Lupas, A. N., C. P. Ponting, et al. (2001). "On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?" J Struct Biol **134**(2-3): 191-203.

Mao, H., S. A. White, et al. (1999). "A novel loop-loop recognition motif in the yeast ribosomal protein L30 autoregulatory RNA complex." Nat Struct Biol **6**(12): 1139-47.

Marchler-Bauer, A., J. B. Anderson, et al. (2005). "CDD: a Conserved Domain Database for protein classification." Nucleic Acids Res **33**(Database issue): D192-6.

Marchler-Bauer, A., A. R. Panchenko, et al. (2002). "CDD: a database of conserved domain alignments with links to domain three-dimensional structure." Nucleic Acids Res **30**(1): 281-3.

Martin, J. L. (1995). "Thioredoxin--a fold for all reasons." Structure **3**(3): 245-50.

Mateu, M. G. and A. R. Fersht (1999). "Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to impaired hetero-oligomerization." Proc Natl Acad Sci U S A **96**(7): 3595-9.

Maxwell, A. (1992). "The molecular basis of quinolone action." J Antimicrob Chemother **30**(4): 409-414.

McGuffin, L. J., K. Bryson, et al. (2000). "The PSIPRED protein structure prediction server." Bioinformatics **16**(4): 404-5.

McNeil, B. J. and J. A. Hanley (1984). "Statistical approaches to the analysis of receiver operating characteristic (ROC) curves." Med Decis Making **4**(2): 137-50.

Mirny, L. A. and M. S. Gelfand (2002). "Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors." J Mol Biol **321**(1): 7-20.

Morais Cabral, J. H., A. P. Jackson, et al. (1997). "Crystal structure of the breakage-reunion domain of DNA gyrase." Nature **388**(6645): 903-6.

Moult, J., K. Fidelis, et al. (2005). "Critical assessment of methods of protein structure prediction (CASP)--round 6." Proteins **61 Suppl 7**: 3-7.

Murzin, A. G. (1998). "How far divergent evolution goes in proteins." Curr Opin Struct Biol **8**(3): 380-7.

Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-40.

Nagano, N., C. A. Orengo, et al. (2002). "One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions." J Mol Biol **321**(5): 741-65.

Nakamura, H., K. Nakamura, et al. (1997). "Redox regulation of cellular activation." Annu Rev Immunol **15**: 351-69.

Nemergut, M., Mizzen CA, Stukenberg T, Allis CD, Macara IG (2001). "Chromatin docking and exchange activity enhancement of RCC1 by histones H2A and H2B." Science **292**: 1540-1543.

Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-217.

Orengo, C. A., J. E. Bray, et al. (2002). "The CATH protein family database: a resource for structural and functional annotation of genomes." Proteomics **2**(1): 11-21.

Orengo, C. A., T. P. Flores, et al. (1993). "Identification and classification of protein fold families." Protein Eng **6**(5): 485-500.

Orengo, C. A., A. D. Michie, et al. (1997). "CATH--a hierarchic classification of protein domain structures." Structure **5**(8): 1093-108.

Ortiz, A. R., C. E. Strauss, et al. (2002). "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison." Protein Sci **11**(11): 2606-21.

Paccanaro, A., J. A. Casbon, et al. (2006). "Spectral clustering of protein sequences." Nucleic Acids Res **34**(5): 1571-80.

Palliser, C. C. and D. A. Parry (2001). "Quantitative comparison of the ability of hydropathy scales to recognize surface beta-strands in proteins." Proteins **42**(2): 243-55.

Palm, G. J., E. Billy, et al. (2000). "Crystal structure of RNA 3'-terminal phosphate cyclase, a ubiquitous enzyme with unusual topology." Structure Fold Des **8**(1): 13-23.

Paoli, M. (2001). "Protein folds propelled by diversity." Prog Biophys Mol Biol **76**(1-2): 103-130.

Pearl, F. M., D. Lee, et al. (2000). "Assigning genomic sequences to CATH." Nucleic Acids Res **28**(1): 277-82.

Pei, J. and N. V. Grishin (2001). "AL2CO: calculation of positional conservation in a protein sequence alignment." Bioinformatics **17**(8): 700-12.

Pei, J. and N. V. Grishin (submitted). "PROMALS: towards accurate multiple sequence alignments of distantly related proteins."

Pollock, D. D. and W. R. Taylor (1997). "Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution." Protein Eng **10**(6): 647-57.

Ponting, C. P. and R. B. Russell (1995). "Swaposins: circular permutations within genes encoding saposin homologues." Trends Biochem Sci **20**(5): 179-80.

Qi, Y., J. Pei, et al. (2002). "C-terminal domain of gyrase A is predicted to have a beta-propeller structure." Proteins **47**(3): 258-64.

Reece, R. J. and A. Maxwell (1991). "The C-terminal domain of the Escherichia coli DNA gyrase A subunit is a DNA-binding protein." Nucleic Acids Res **19**(7): 1399-1405.

Renault, L., J. Kuhlmann, et al. (2001). "Structural basis for guanine nucleotide exchange on Ran by the regulator of chromosome condensation (RCC1)." Cell **105**(2): 245-55.

Renault, L., N. Nassar, et al. (1998). "The 1.7 A crystal structure of the regulator of chromosome condensation (RCC1) reveals a seven-bladed propeller." Nature **392**(6671): 97-101.

Robinson, A. B. and L. R. Robinson (1991). "Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins." Proc Natl Acad Sci U S A **88**(20): 8880-4.

Rost, B. (1995). "TOPITS: threading one-dimensional predictions into three-dimensional structures." Proc Int Conf Intell Syst Mol Biol **3**: 314-21.

Rost, B., R. Schneider, et al. (1997). "Protein fold recognition by prediction-based threading." J Mol Biol **270**(3): 471-80.

Rychlewski, L. and D. Fischer (2005). "LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction." Protein Sci **14**(1): 240-5.

Rychlewski, L., L. Jaroszewski, et al. (2000). "Comparison of sequence profiles. Strategies for structural predictions using sequence information." Protein Sci **9**(2): 232-41.

Rychlewski, L., B. Zhang, et al. (1998). "Fold and function predictions for Mycoplasma genitalium proteins." Fold Des **3**(4): 229-38.

Sadreyev, R. and N. Grishin (2003). "COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance." J Mol Biol **326**(1): 317-36.

Salem, G. M., E. G. Hutchinson, et al. (1999). "Correlation of observed fold frequency with the occurrence of local structural motifs1." Journal of Molecular Biology **287**(5): 969-981.

Saraf, M. C., G. L. Moore, et al. (2003). "Using multiple sequence correlation analysis to characterize functionally important protein regions." Protein Eng **16**(6): 397-406.

Schaffer, A. A., L. Aravind, et al. (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Res **29**(14): 2994-3005.

Schaffer, A. A., Y. I. Wolf, et al. (1999). "IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices." Bioinformatics **15**(12): 1000-11.

Shi, J., T. L. Blundell, et al. (2001). "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties." J Mol Biol **310**(1): 243-57.

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

Siebold, C., L. F. Garcia-Alles, et al. (2003). "A mechanism of covalent substrate binding in the x-ray structure of subunit K of the Escherichia coli dihydroxyacetone kinase." Proc Natl Acad Sci U S A **100**(14): 8188-92.

Siew, N., A. Elofsson, et al. (2000). "MaxSub: an automated measure for the assessment of protein structure prediction quality." Bioinformatics **16**(9): 776-85.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-60.

Suel, G. M., S. W. Lockless, et al. (2003). "Evolutionarily conserved networks of residues mediate allosteric communication in proteins." Nat Struct Biol **10**(1): 59-69.

Sunyaev, S. R., F. Eisenhaber, et al. (1999). "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations." Protein Eng **12**(5): 387-94.

Tang, C. L., L. Xie, et al. (2003). "On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles." J Mol Biol **334**(5): 1043-62.

Tatusov, R. L., S. F. Altschul, et al. (1994). "Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks." Proc Natl Acad Sci U S A **91**(25): 12091-5.

Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Teodorescu, O., T. Galor, et al. (2004). "Enriching the sequence substitution matrix by structural information." Proteins **54**(1): 41-8.

Tete-Favier, F., D. Cobessi, et al. (2000). "Crystal structure of the Escherichia coli peptide methionine sulphoxide reductase at 1.9 A resolution." Structure Fold Des **8**(11): 1167-78.

Tillier, E. R. and T. W. Lui (2003). "Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments." Bioinformatics **19**(6): 750-5.

Tress, M., C. H. Tai, et al. (2005). "Domain definition and target classification for CASP6." Proteins **61 Suppl 7**: 8-18.

Walker, D. R. and E. V. Koonin (1997). "SEALS: a system for easy analysis of lots of sequences." Ismb **5**: 333-339.

Wang, J. C. (1996). "DNA topoisomerases." Annu Rev Biochem **65**: 635-692.

Weichsel, A., J. R. Gasdaska, et al. (1996). "Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer." Structure **4**(6): 735-51.

Westbrook, J., Z. Feng, et al. (2003). "The Protein Data Bank and structural genomics." Nucleic Acids Res **31**(1): 489-91.

Wikoff, W. R., L. Liljas, et al. (2000). "Topologically linked protein rings in the bacteriophage HK97 capsid." Science **289**(5487): 2129-33.

Wolan, D. W., S. E. Greasley, et al. (2002). "Structural insights into the avian AICAR transformylase mechanism." Biochemistry **41**(52): 15505-13.

Wolf, Y. I., S. E. Brenner, et al. (1999). "Distribution of protein folds in the three superkingdoms of life." Genome Res **9**(1): 17-26.

Xiang, S., S. A. Short, et al. (1996). "Cytidine deaminase complexed to 3-deazacytidine: a "valence buffer" in zinc enzyme catalysis." Biochemistry **35**(5): 1335-41.

Xie, L. and P. E. Bourne (2005). "Functional Coverage of the Human Genome by Existing Structures, Structural Genomics Targets, and Homology Models." PLoS Comput Biol **1**(3): e31.

Yamawaki, H., J. Haendeler, et al. (2003). "Thioredoxin: a key regulator of cardiovascular homeostasis." Circ Res **93**(11): 1029-33.

Yona, G. and M. Levitt (2002). "Within the twilight zone: a sensitive profile-profile comparison tool based on information theory." J Mol Biol **315**(5): 1257-75.

Zechiedrich, E. L., A. B. Khodursky, et al. (2000). "Roles of topoisomerases in maintaining steady-state DNA supercoiling in Escherichia coli." J Biol Chem **275**(11): 8103-8113.

Zemla, A. (2003). "LGA: A method for finding 3D similarities in protein structures." Nucleic Acids Res **31**(13): 3370-4.

Zemla, A., C. Venclovas, et al. (1999). "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment." Proteins **34**(2): 220-3.

Zhang, Y. and J. Skolnick (2004). "Scoring function for automated assessment of protein structure template quality." Proteins **57**(4): 702-10.

Zhu, J. and Z. Weng (2005). "FAST: a novel protein structure alignment algorithm." Proteins **58**(3): 618-27.

VITAE

Yuan Qi was born in Qingdao, Shandong Province, P. R. China, on June 4, 1974, daughter of Rusong Qi and Ruihua Li. In 1993, she graduated as the top one student from the 19th High School, Qingdao, Shandong and was admitted to the University of Science of Technology of China (USTC), Hefei, with the entrance examinations waived due to her excellent academic records. She received the degree of Bachelor of Science with a major in cell biology and biophysics from USTC in July 1998. Right after her graduation from college, she came to the U.S. to pursue further education and was awarded the degree of Master of Arts majoring in Biochemistry from the University of Scranton, Pennsylvania, in May 2000. She entered the Graduate school of Biomedical Sciences at the University of Texas Southwestern Medical Center at Dallas in August 2000 and joined the Computational Biology lab of Dr. Nick Grishin in 2001.

Permanent Address: 34 Binxian Street, #502
Qingdao, Shandong 266021
P. R. China