EXPLORING SEQUENCE-STRUCTURE-FUNCTION RELATIONSHIPS IN PROTEINS USING CLASSIFICATION SCHEMES

APPROVED BY SUPERVISORY COMMITTEE

Nick V. Grishin, Ph. D.; Advisor

Stephen R. Sprang, Ph. D.; Committee Chair

Alexander Pertsemlidis, Ph. D.

Zbyszek Otwinowski, Ph. D.

To Mike, Jan, Andy, and Barb Cheek.

In honor of my Lord Jesus Christ.

EXPLORING SEQUENCE-STRUCTURE-FUNCTION RELATIONSHIPS IN PROTEINS USING CLASSIFICATION SCHEMES

by

SARA ANNE CHEEK

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, 2005

EXPLORING SEQUENCE-STRUCTURE-FUNCTION RELATIONSHIPS IN PROTEINS USING CLASSIFICATION SCHEMES

Publication No. _____

Sara Anne Cheek, Ph. D.

The University of Texas Southwestern Medical Center at Dallas, 2005

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

With the rapid growth in the number of available protein sequences and structures, the necessity of interpreting this data in comprehensive and meaningful ways becomes increasingly apparent. Identifying and categorizing the functional, structural, and evolutionary relationships between proteins is a key step in understanding protein evolution. Protein classification is a useful means of organizing biological data for the purpose of exploring these sequence-structure-function relationships in proteins. In this work, two-tier classification schemes are constructed for the organization of large protein classes. One level of this hierarchy reflects structural similarity ("fold groups"), while the second level indicates an evolutionary relationship between members ("families").

Kinases are a ubiquitous group of enzymes that participate in a variety of cellular pathways. Despite that all kinases catalyze similar phosphoryl transfer reactions, they display remarkable diversity in structural fold and substrate specificity. All available kinase sequences and structures have been classified into fold groups and families. This classification presents the first comprehensive structural annotation of a large functional class of proteins. The question of how different structural folds accomplish the same fundamental elements of the kinase reaction is investigated.

Disulfide-rich domains are small protein domains whose global folds are stabilized predominantly by disulfide bonds. In order to understand the structural and functional diversity among available disulfide-rich proteins, a comprehensive classification of these domains has been performed. The resulting fold groups and families describe more distant structural and evolutionary relationships than previously acknowledged among disulfide-rich domains. Variations in disulfide bonding patterns of these domains are also evaluated.

Several existing classification databases have been developed for the purpose of cataloguing all available protein structures. Because such databases are often manually curated, recently solved structures are not included and useful information regarding their relatedness to other proteins is not immediately available. To address this limitation, an algorithm has been developed to make classification assignments with evolutionary relevance for domains in newly solved structures, with the objective of reliably reproducing assignments to an existing classification scheme in an automatic manner.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Cheek, S, Krishna, SS, and Grishin, NV. "Structural Classification of Small Disulfide-Rich Protein Domains." *Submitted*.

Cheek, S, Ginalski, K, Zhang, H, and Grishin, NV. (2005) "A comprehensive update of the sequence and structure classification of kinases." *BMC Struct Biol* **5**(1):6.

Kinch, LN, Cheek, S, and Grishin, NV. (2005) "EDD, a novel phosphotransferase domain common to mannose transporter EIIA, dihydroxyacetone kinase, and DegV." *Protein Sci* **14**(2):360-7.

Cheek, S, Qi,Y, Krishna, SS, Kinch, LN, and Grishin, NV. (2004) "SCOPmap: automated assignment of protein structures to evolutionary superfamilies." *BMC Bioinformatics* **5**(1):197.

Cheek, S, Zhang, H, and Grishin, NV. (2002) "Sequence and structure classification of kinases." *J Mol Biol* **320**(4):855-81.

Zhang, H, Zhou, T, Kurnasov, O, Cheek, S, Grishin, NV, and Osterman, A. (2002) "Crystal structures of *E. coli* nicotinate mononucleotide adenylyltransferase and its complex with deamido-NAD." *Structure* **10**(1):69-79.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIR synthetase – aminoimidazole ribonucleotide synthetase

AL2CO – tool for conservation analysis of a given multiple sequence alignment

ASKHA superfamily– acetate and sugar kinase/hsc70/actin superfamily

BLAST – Basic Local Alignment Search Tool

BPTI – bovine pancreatic trypsin inhibitor

BBI – Bowman-Birk inhibitor

BI-VI – bromelain inhibitor VI

BLOSUM – BLOCKS substitution matrix

CATH – classification database (class, architecture, topology, homologous superfamily)

CAZy – carbohydrate-active enzymes database

CBDx – cellulose binding domain of xylanase A

CDART – Conserved Domain Architecture Retrieval Tool

CDD – Conserved Domain Database

CDDe – cellulose docking domain of endoglucanase Cel45A

COG – Clusters of Orthologous Groups

COMPASS –Comparison of Multiple Protein Alignments with Assessment of Statistical Significance

CRISP family – cysteine-rich secretory protein family

DhaK – dihydroxyacetone kinase

DHS – Dictionary of Homologous Superfamilies

DnaProt – DNA-binding proteins database

Dol-P – dolichyl monophosphate

D_Z – Z-score calculated by DaliLite

EC – Enzyme Commission

EGF – epidermal growth factor

F2CS – FSSP to CATH and SCOP prediction server

FSSP – Families of Structurally Similar Proteins

FunCat – Functional Catalogue database

gi – NCBI gene identification number

GO – Gene Ontology

GHL family – DNA gyrase/Hsp90/MutL family

GHMP – galactokinase/homoserine kinase/mevalonate kinase/phosphomevalonate kinase

GRDB – Gene Relational Database system

HAD-like – L-2-haloacid dehalogenase-like

HK – histidine kinase

HMM – hidden Markov model

HPrK/P – histidine-containing phosphocarrier protein kinase/phosphatase

HPPK – amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase

I3P3K – inositol 1,4,5-trisphosphate 3-kinase

I5P2K – inositol 1,3,4,5,6-pentakisphosphate 2-kinase

Ig – immunoglobulin

IGFBP – insulin-like growth factor binding protein

IP – inositol polyphosphate

KOG – eukaryotic orthologous groups

MAMMOTH – Matching Molecular Models Obtained from Theory

M_Z – Z-score calculated by MAMMOTH

NCBI – National Center for Biotechnology Information

NDP kinase – nucleoside diphosphate kinase

nr – non-redundant protein database

PALI – phylogeny and alignment of homologous protein structures

PANTHER – Protein Analysis Through Evolutionary Relationships database

PCMA – profile consistency multiple sequence alignment

PDB – Protein Data Bank

PEPCK – phosphoenolpyruvate carboxykinase

Pfam – Protein family database

PIPK – type IIβ phosphatidyl*i*nositol phosphate kinase

PIRSF – PIR (Protein Information Resource) superfamilies database

PK – protein kinase

PKD – polycystic kidney disease

PLD – phospholipase D

P-loop – phosphate-binding loop

PMK – phosphomevalonate kinase

PPK – polyphosphate kinase

PRODISTIN – protein distance based on interactions

PSI-BLAST – Position-Specific Iterated BLAST

RFK – riboflavin kinase

RMSD – root-mean-square deviation

RPS-BLAST – Reverse Position-Specific BLAST

SAM domain – sterile alpha motif domain

SCOP – Structural Classification of Proteins database

SEALS – System for Easy Analysis of Lots of Sequences

SK – shikimate kinase

SMART – Simple Modular Architecture Research Tool

SMB – somatomedin B

SSAP – Sequential Structure Alignment Program

SSM – Secondary Structure Matching

SUPFAM – SuperFamily Database

T6P kinase – tagatose 6-phosphate kinase

TAP – tick anticoagulant protein

TIGRFAMs – TIGR (The Institute for Genomic Research) protein families database

TPK – thiamin pyrophosphokinase

TSP – thrombospondin

UMP/CMP kinase – pyrimidine nucleoside (uridine or cytidine) monophosphate kinase

Undec-P – undecaprenyl monophosphate

UniProt – Universal Protein Resource

ωTL – ω-conotoxin-like

CHAPTER 1:
General Introduction

Recent years have seen an explosion in the amount of available protein sequence and structure data. With this rapid growth comes the necessity of interpreting this information in comprehensive and meaningful ways. Protein classification is a tool that is particularly well suited to this task. Grouping proteins based on shared characteristics, whether functional, structural, or evolutionary in nature, can achieve a several-fold reduction in the data such that a large set of proteins can be described by a small number of representatives. Furthermore, protein classification is a useful means of studying various aspects of sequence, structural, and functional similarities within and between protein families. In this dissertation, classification of large protein classes is carried out in order to further our understanding of sequence-structure-function relationships in proteins. More specifically, this work involves the identification of previously undetected evolutionary links, the investigation of how specific attributes of function are manifest in structural folds, and the evaluation of functional/structural convergence and divergence within particular classes of proteins.

1.1 PROTEIN CLASSIFICATION

The general logic of classification is to simplify some complex set of data by grouping together those entities that share common attributes. The most well known biological classification addresses the categorization of living organisms (i.e. phylogenetic taxonomy) and is commonly used in investigating the theory of species evolution. Similarly, the application of classification to the protein universe is a convenient tool for the study of molecular evolution. More specifically, protein classification enables the analysis of characteristics such as sequence, structural, and functional similarity within and between protein families. The utility of protein

classification in the study of sequence-structure-function relationships in proteins is reviewed in section 1.2.

Existing protein classification schemes are generally organized based on relationships among structures, sequences, and/or functions. The advantages and disadvantages of each of these three main approaches are described in the following sections.

1.1.1 Sequence-based classification databases

Databases that group proteins according to sequence similarity (i.e. homology) are a popular type of protein classification scheme. These databases are typically not hierarchical in nature, unlike most structure-based classification databases (see section 1.1.2). However, the general purpose of these schemes is to reduce the complexity of the available protein dataset and so the term "classification" is not a misnomer. A key advantage of sequence-based classification is that the protein groupings are not limited to data from solved structures. Since it is estimated that a structural representative is currently available for only 20% of known protein families (Wolf, Grishin et al. 2000; Yan and Moult 2005), homology-based classification schemes may encompass 5-fold more protein information than structure-based schemes. However, very distant evolutionary relationships are often difficult to detect in the absence of structural information, and many presumably unrelated protein families identified by sequence-based classification schemes will likely be merged in the future. In a structure-based classification scheme, such potential links can be recognized by searching among the broader levels of the hierarchy (see section 1.1.2). The inability to scrutinize these speculative connections in the non-hierarchical sequence-based classification databases is a fundamental drawback in using these tools.

Sequence- or homology-based classification databases are commonly used to study the homologs of a particular protein of interest or to analyze protein families as a whole (for example, to investigate sequence or functional variation among evolutionary

neighbors). A large number of sequence-based protein classifications are currently available. A few popular examples include Pfam, COG, and SMART.

Pfam

The Pfam (protein family) database of is a collection of multiple sequence alignments of protein domains (Bateman, Coin et al. 2004). The philosophy of Pfam is to combine automated methods with visual inspection in order to capitalize on both the comprehensiveness of automatic approaches and the higher quality of manual curation (Sonnhammer, Eddy et al. 1997). Pfam consists of two sections: A and B.

Pfam-A contains high-quality alignments for well-characterized families. In Pfam-A, the following components are provided for each protein domain family: a description including any available functional and structural information, salient literature references, cross-links to other databases (such as SCOP (Murzin, Brenner et al. 1995), PRINTS (Attwood, Beck et al. 1994), InterPro (Apweiler, Attwood et al. 2000), and the PDB (Berman, Westbrook et al. 2000)), visualization of the domain architecture of proteins containing the domain in question, a phylogenetic tree, two high-quality alignments (seed and full), and the hidden Markov model (HMM) describing the family. The two alignments and the HMM for each family are the core of the Pfam-A database. The seed alignment undergoes careful manual curation when a protein family is initially added to the Pfam-A database and is rarely adjusted after that point. This seed alignment is then used to generate an HMM describing the protein domain family in question. The HMM is subsequently used to build the full alignment, which includes all of the sequences assigned to that Pfam-A family.

Pfam-B includes sequence families automatically generated by clustering any remaining (i.e. unassigned in Pfam-A) proteins from a non-redundant set derived from ProDom (Sonnhammer and Kahn 1994). The Pfam-B groupings and alignments are generally of lower quality, but can be helpful in cases when no domains annotated in Pfam-A are detected in a particular protein sequence.

Because all members of a Pfam family must be described by a single HMM, the database is fairly conservative with, in some cases, divergent members of the same protein family represented by separate Pfam families. More remote evolutionary links between different Pfam families are acknowledged in Pfam clans. Clans are usually established based on the presence of common sequence motifs or the similarity of structural folds. Currently, over 150 Pfam clans linking two or more Pfam families have been identified.

Pfam offers an interactive web server, which can assign putative Pfam assignments to a query sequence. This tool uses the HMMs describing the Pfam families to analyze the potential Pfam classification of domains within the query protein.

The Pfam database is generally updated every 3-6 months. The most recent version of Pfam-A (v18.0; released August 2005) classifies 1,426,410 protein sequences into 7973 families. Pfam-B currently includes 128,469 clusters encompassing 327,279 protein sequences.

COG and KOG

COG (Clusters of Orthologous Groups) (Tatusov, Koonin et al. 1997) is a classification of protein domains from the completed genomes of prokaryotes and unicellular eukaryotes. COGs are constructed by grouping together proteins that correspond to the best inter-genome hits found by BLAST. Each COG contains proteins from at least 3 different species and therefore describes an ancient conserved domain. Similarly, KOGs (eukaryotic Orthologous Groups) are groups of protein domains from the completed genomes of eukaryotic species (Tatusov, Fedorova et al. 2003).

For each cluster described in the COG/KOG databases, the phylogenetic distribution of proteins in that COG/KOG as well as a graphical view of significant hits (in terms of sequence similarity) between members is provided. Additionally, for each protein member of the COGs/KOGs, a graphical view of BLAST hits with detected COG/KOG domains is available.

Query sequences can be classified in existing COGs or KOGs by the COGnitor and KOGnitor tools, respectively.

Currently, COG classifies 129,326 proteins from 66 unicellular genomes (prokaryotic and eukaryotic) into 4873 clusters, and KOG classifies 60,758 proteins from 7 genomes into 4852 clusters.

SMART

Similar to Pfam, SMART (Simple Modular Architecture Research Tool) (Letunic, Copley et al. 2004) utilizes HMMs to describe conserved domains within proteins. However, unlike Pfam, which is not limited to any specific functional class or taxonomy, the original intention of SMART was to provide a tool for the study of the evolution of function in multi-domain proteins (Schultz, Milpetz et al. 1998). Although SMART has expanded somewhat beyond this initial motivation, this database still heavily emphasizes domain families from nuclear, signaling, and extracellular proteins in eukaryotic organisms.

For each SMART domain family, annotation summarizing available information about function (in terms of cellular role, molecular interaction, involvement in human disease, and identification of functionally important residues), subcellular localization, phylogenetic distribution, and structural fold is provided. Alignments of domain family members, significant literature references, and links to other protein databases are also available. SMART domains identified in non-redundant protein databases are recorded in a relational database system. There are two modes of SMART which are differentiated by the underlying protein database: normal SMART includes all proteins found in databases such as SwissProt, while genomic SMART includes only proteins from completely sequenced genomes.

SMART provides a web interface that allows users to submit query sequences and search for SMART domains (identified by searching with HMMs) as well as intrinsic sequence features such as transmembrane regions, disordered regions, coiled-coils, signal peptides, and internal repeats. A second option involves automated searches through

protein databases (in which SMART domains are already identified) in order to detect user-defined combinations of specific domain architectures, functional terms, and taxonomies.

Currently, the SMART database (v4.1) includes 678 domain families.

Sequence motif-based classification

The previous examples are classification schemes dedicated to identifying and organizing sequence similarity between entire proteins or domains. Another approach to homology-based classification entails the recognition of conserved sequence motifs that are characteristic of a particular protein family, rather than searching for sequence similarity over the entire length of a protein or domain. This approach is based on the observation that certain parts of proteins are more important than others due to functional or structural requirements. As a result of evolutionary constraints on these regions, certain "signature motifs" tend to diverge much slower than the rest of the protein sequence and are therefore useful for identifying family members in the absence of overall sequence similarity. Two examples of classification schemes based on the detection of sequence motifs are the PROSITE and PRINTS databases.

The PROSITE database (Sigrist, Cerutti et al. 2002) characterizes protein families using both patterns describing short motifs and profiles describing entire proteins or domains. Sequence motifs in the PROSITE database are defined by patterns that explicitly describe the length of the motif and the amino acid variability allowed at each site. For example, the Walker-A nucleotide-binding P-loop motif (Walker, Saraste et al. 1982) is described by the following pattern in PROSITE: [AG] - x(4) - G - K - [ST]. Typical PROSITE patterns are short and highly specific and usually correspond to biologically important regions such as enzyme catalytic sites, substrate-binding sites, residues involved in coordinating metal ions, etc. This database initially consisted only of patterns, although PROSITE now addresses more divergent sequence similarities by including profiles that cover the entire length of protein families. The ScanProsite tool

(Gattiker, Gasteiger et al. 2002) is able to search for PROSITE patterns within a query sequence or to search for sequences containing a query pattern.

PRINTS (Attwood, Beck et al. 1994) is a database of protein fingerprints. Unlike PROSITE, which typically characterizes short, functionally-relevant regions using a single pattern, a fingerprint in the PRINTS database is defined as a set of conserved sequence motifs that typify a protein family. These fingerprints are crafted by excising conserved regions within multiple sequence alignments. The FingerPRINTScan tool (Scordis, Flower et al. 1999) can be used to identify these fingerprints within a query sequence.

Integrated databases of protein families

Some classification resources combine information about protein families from several independent databases. Because each member database is constructed based on slightly (or, in some cases, substantially) different methodology, each contributes different advantages and disadvantages, making their integration a valuable approach. Two of the most popularly used of these integrated databases are CDD and InterPro.

CDD (Conserved Domain Database) (Marchler-Bauer, Anderson et al. 2005) is a collection of domain alignments for families represented in Pfam, COG, and SMART. Also included in CDD are additional alignments for ancient conserved domains that are identified by phylogenetic analysis. CDD domains within query proteins can be detected using RPS-BLAST (Marchler-Bauer, Anderson et al. 2003). Another tool associated with CDD is CDART (Conserved Domain Architecture Retrieval Tool) (Geer, Domrachev et al. 2002), which can be used to identify proteins containing a similar set of domains as those found in a query sequence.

InterPro (Apweiler, Attwood et al. 2000) is another resource that integrates protein family information from multiple diverse sources. Several HMM-based resources are included: Pfam (Bateman, Coin et al. 2004), SMART (Letunic, Copley et al. 2004), SUPERFAMILY (Gough, Karplus et al. 2001), TIGRFAMs (Haft, Loftus et al. 2001), PIRSF (Wu, Nikolskaya et al. 2004), PANTHER (Thomas, Kejariwal et al. 2003), and

Gene3D (Buchan, Rison et al. 2003). Another member database, ProDom (Sonnhammer and Kahn 1994), establishes protein families based on PSI-BLAST hits. PROSITE (Sigrist, Cerutti et al. 2002) and PRINTS (Attwood, Beck et al. 1994), which focus on conserved sequence motifs rather than entire domains or proteins, are incorporated as well. Protein signatures (i.e. families, domains, or motifs) described by the member databases can be detected in query sequences by use of the web-based InterProScan tool (Zdobnov and Apweiler 2001).

1.1.2 Structure-based classification databases

Several structural classification databases have been developed for the purpose of cataloging all available protein structures. These classification schemes are popularly used for identifying the structural neighbors of a protein of interest. In some cases, these databases also incorporate level(s) of homology into the classification hierarchy, so that structurally similar proteins are also arranged according to their evolutionary relatedness. Because homologous proteins often maintain structural similarity even when no significant sequence similarity can be detected, structural classification schemes can potentially reveal remote evolutionary links that might otherwise go unrecognized. One critical limitation in the use of structural classification schemes concerns the availability of structural data, which is currently far less abundant than protein sequence data. Nonetheless, structural classification tools are generally considered to be the most informative for the purpose of studying sequence-structure-function relationships in proteins. Among the most popularly used structural classification databases are SCOP, CATH, and Dali Domain Dictionary.

SCOP

The SCOP database (Structural Classification of Proteins) (Murzin, Brenner et al. 1995) provides a comprehensive classification of nearly all proteins with known structure, based on their structural and evolutionary relationships. The SCOP hierarchy

consists of six levels. The first two levels organize domains by their structural features. These levels are *class*, which refers primarily to secondary structure make-up (e.g. "all-alpha"), and *fold*, which signifies that proteins possess the same core secondary structure elements with both the same spatial arrangement (architecture) and same connections (topology). The next two levels reflect evolutionary relationships, with the *superfamily* level denoting more remote evolutionary connections and the *family* level identifying close homologs that often perform related functions. The final two levels specify the exact *protein* and *species* to which the domains belong. Thus, while SCOP is a classification of protein structures, the hierarchy also acknowledges evolutionary relatedness among domains.

The SCOP classification was constructed and is maintained/updated using a combination of visual inspection and automatic protein comparison tools. Because SCOP relies largely on manual curation, it is considered by many to be the gold standard of structural classification databases. Unfortunately, this dependency on manual intervention also results in a substantial drawback; the availability of SCOP assignments generally lags many months behind the PDB database. For example, in October 2005, the most current version of the SCOP database (v1.69, released July 2005) contained only those structures that were found in the PDB prior to October 1, 2004. Therefore, only a few months after the most recent update, SCOP was already outdated by one year and roughly 5500 structures. Clearly, this poses a problem in the use of SCOP for the study of recently solved protein structures.

In order to partially address this shortcoming, SCOP is now associated with the protein structure comparison tool SSM (Secondary Structure Matching) (Krissinel and Henrick 2004). SSM identifies potential structural neighbors for a given query structure. The SCOP classification of these structural neighbors is identified in the output, so the SCOP assignment of the query structure can, in some cases, be inferred from the SSM results. However, SSM does not predict a unique position within the SCOP hierarchy, and determining the appropriate SCOP classification for a query structure consequently requires a considerable amount of manual study in non-trivial cases. For example, the

structure of hypothetical protein Pg0816 from *Porphyromonas gingivalis* (Chang, Quartey et al. *To be published*) was recently solved (pdb|2apl; released September 27, 2005) and is not currently classified in the SCOP database. This 157-residue protein adopts an entirely α-helical fold. SSM finds that the structure of Pg0816 does not closely resemble any other protein previously classified by SCOP (i.e. no hits with Z-score > 2.5), and instead suggests 102 potential structural neighbors from 10 different SCOP folds. Thus, establishing where hypothetical protein Pg0816 should be assigned within the SCOP hierarchy would entail careful manual analysis of many other protein structures. Another restriction is that the predictions made by SSM do not necessarily reflect homology relationships because the algorithm considers only the structural similarities between proteins. Determining the evolutionary implications of the SSM predictions requires additional study of the proteins in question. These limitations of the otherwise invaluable SCOP database are addressed in Chapter 4.

SCOP also links to other databases that assist users in the further study of SCOP domains by presenting, for example, structure-based pairwise and multiple alignments of SCOP families (PALI; Phylogeny and Alignment of homologous protein structures) (Balaji, Sujatha et al. 2001) or the expansion of SCOP families with the identification of remote homologs (SUPFAM) (Pandit, Gosar et al. 2002).

The SCOP database is generally updated on a biannual basis. The most recent version of SCOP (v1.69, released July 2005) classifies 70,859 protein domains into 945 folds and 1539 superfamilies.

CATH

CATH (Orengo, Michie et al. 1997) is another classification of protein structures that organizes domains according to both structural and evolutionary criteria. CATH was named after the first four levels of its 7-level hierarchy. The first level, *class* (C), describes the secondary structure composition of the protein domain (e.g. “mainly-alpha”). The second level, *architecture* (A), refers to the general spatial arrangement of the secondary structure elements without regard for topology (e.g., “sandwich” or

"horseshoe"). Domains are categorized by the connectivity of their secondary structure elements at the third level, *topology* (T). The fourth level, *homologous superfamily* (H), clusters protein domains that are evolutionarily related. *Sequence families* (S) form the fifth level and include homologous protein domains with highly similar sequences (>35% identity). The final two levels group *non-identical* (N) and *identical* (I) domains based on shared sequence identity of ≥95% and 100%, respectively. CATH is updated largely by the use of automatic sequence and structure comparison tools, although some manual evaluation is also performed, most notably at the architecture level.

CATH provides a server that attempts to automatically classify structures by identifying sequence similarity (from searching against the CATH-HMM library) and structural similarity (using structure comparison tools SSAP (Taylor and Orengo 1989) and CATHEDRAL (Harrison, Pearl et al. 2003)) between the query and known CATH domains. This tool is designed to make putative assignments at the homologous superfamily level and assess the statistical significance of those links. CATH is also associated with DHS (Dictionary of Homologous Superfamilies), which provides additional information, such as multiple alignments and functional annotations, for each homologous superfamily (Bray, Todd et al. 2000).

The CATH database is usually updated once a year. The most recent version of CATH (v2.6.0, released April 2005) classifies 67,054 protein domains into 907 topologies and 1572 homologous superfamilies.

Dali Domain Dictionary

The Dali Domain Dictionary (Holm and Sander 1998) is another scheme that classifies protein structures based on structural and evolutionary considerations. This classification is a supplement of the FSSP database (Families of Structurally Similar Proteins) (Holm and Sander 1994), which clusters and structurally aligns proteins based on all-against-all comparisons performed by the Dali structure comparison tool (Holm and Sander 1995). While the FSSP database merely provides non-hierarchical groupings of proteins based on significant structural similarity, the Dali Domain Dictionary

classifies domain structures in a 4-level hierarchy meant to acknowledge both structural and evolutionary relatedness.

The first level of the Dali Domain Dictionary, *fold space attractor region*, describes secondary structure composition and, in some cases, a very general topological category (e.g. "all-alpha" or "antiparallel beta-barrels"). There are currently 5 defined attractors, with two additional attractors that contain structures which fall somewhere in between the 5 established categories and structures that are currently unique. *Globular folding topology*, or fold type, is the second level and groups together domains with architectural and topological similarities. Within fold types, average pairwise Dali Z-scores between members are greater than 2. Level three, *functional family*, suggests potential evolutionary relationships based on functional or sequence similarity in addition to significant structural similarity. These speculative homology links are established using a neural network strategy that considers sequence and functional information from various sources, such as PSI-BLAST (Altschul, Madden et al. 1997) and UniProt (Apweiler, Bairoch et al. 2004). The fourth level, *sequence family*, includes homologous domain representatives with greater than 25% sequence identity. Classification in the Dali Domain Dictionary is an entirely automated process (Dietmann and Holm 2001). However, this database is not currently associated with any interactive tools that suggest potential Dali Domain Dictionary classification assignments for given query structures.

The most recent version of the Dali Domain Dictionary (v.3.1beta; released March 2001) classifies 35,492 protein domains into 1088 fold types and 2073 functional families.

1.1.3 Functional classification

Function-based classification schemes

Classification schemes based on functional similarities are fundamentally different than those based on structural or evolutionary relatedness. First, the idea of functional relatedness is not clearly defined, unlike sequence or structural similarity,

which can be assessed statistically. For example, is pyruvate kinase more similar to hexokinase because these two enzymes both perform similar phosphotransfer reactions, or to pyruvate dehydrogenase because these two enzymes both act on the same substrate, or to 2-phosphoglycerate enolase because these two enzymes are both involved in the gluconeogenesis pathway? Also, it must be clarified whether the term "protein function" refers to a molecular, cellular, or physiological role. To further complicate matters, numerous proteins have multiple functions: they are components of several cellular pathways, or possess the capability of binding to or acting on several different substrates. Classification of these cases would require either that one particular function be selected as predominant over the rest, or that proteins be given multiple (i.e. non-unique) assignments within the classification hierarchy. Lastly, while the association between homology and structural similarity is relatively constant (i.e. homologs usually have related structures, although related structures are not necessarily homologous), their relationship with function is much more variable. For example, homologous proteins may or may not have the same function, and vice versa. Likewise, structural similarity does not typically correspond to functional similarity. Thus, while function is generally considered to be the most important attribute of a protein, it is the most problematic characteristic upon which to base a classification scheme.

Considering these complications, it is perhaps not surprising that few global, function-based classification schemes are currently available. The GO (Gene Ontology) database classifies proteins according to three separate ontologies: molecular functions, biological processes, and cellular components (Ashburner, Ball et al. 2000). Other functional classification systems organize data based on protein-protein interactions (PRODISTIN) (Brun, Chevenet et al. 2003) or involvement in pathways (FunCat) (Ruepp, Zollner et al. 2004). The EC (Enzyme Commission) database classifies enzymes in a hierarchy of class (type of catalytic reaction, such as "transferase"), subclass (often referring to the chemical nature of the donor substrate, such as "transferring a P-containing group"), and sub-subclass (often referring to the chemical nature of the acceptor substrate, such as "alcohol group as acceptor") (Barrett, Canter et al. 1992). The

EC database is essentially a classification of nomenclatures and does not incorporate any evolutionary or structural information.

Classification of specific functional classes

There are, however, several classification schemes that have been developed for the categorization of known members of a particular function. Such databases can be thought of as the application of structure/sequence classification to a particular functional class. Examples include the MEROPS database that catalogs peptidases and peptidase inhibitors (Rawlings, Tolle et al. 2004), the DnaProt database that classifies DNA-binding proteins (Karmirantzou and Hamodrakas 2001), and the CAZy database that describes families of domains within enzymes that create, alter, or break down glycosidic bonds (http://afmb.cnrs-mrs.fr/CAZY/).

1.2 SEQUENCE-STRUCTURE-FUNCTION RELATIONSHIPS IN PROTEINS

Proteins are, in general, evaluated in terms of three fundamental aspects: their sequences or evolutionary relationships, their structures, and their functions. However, it is not well understood how these three features relate to each other. On one hand, global sequence similarity has been demonstrated to be a good indicator structural similarity; pairs of proteins with sequence identity greater than ~25% are generally assumed to adopt similar folds. However, the physical principles that guide sequences to adopt specific global folds have not been resolved, and the challenge of predicting a structural fold based solely on a protein sequence (in the absence of homology) remains problematic (Kryshtafovych, Venclovas et al. 2005). Furthermore, a few examples of fold variation among known homologs have been observed, particularly in small domains (Borden and Freemont 1996; Lauber, Schulz et al. 2003). The relationship between protein function and sequence (or structure) is also inconsistent. In many cases, such as orthologous proteins, sequence similarity does in fact indicate conservation of function.

However, even among close homologs, the functional promiscuity of proteins has been well documented. The classic example is lactate dehydrogenase, which is known to perform both structural and enzymatic roles (Wistow and Piatigorsky 1987). Thus, while the basis of such relationships can be generalized as "significantly similar sequences usually adopt similar structures and often perform related functions", there is still a great deal to be revealed about the nature of sequence-structure-function relationships in proteins.

Therefore, protein classification is a valuable tool because it not only enables the examination of sequence, structural, and functional diversity among classes of related proteins, but it also provides an expedient basis for the study of relationships *between* a protein's (or protein family's) sequence, structure, and function. General applications for the use of protein classification schemes are described below.

1.2.1 Protein family descriptions

The most fundamental result of protein classification is that domains are grouped together with their homologs and/or their structural neighbors. One consequence is that entire protein families can be studied as a whole, which can reveal particular sequence motifs or structural features that are signatures of a given family. For example, Bork *et al* identified five conserved motifs (PHOSPHATE 1, PHOSPHATE 2, ADENOSINE, CONNECT 1, and CONNECT 2) common to the members of a diverse family containing sugar kinases, actin, and heat shock proteins (Bork, Sander et al. 1992). Conversely, sequence differences among close homologs can be used to study determinants of functional variations such as substrate specificity. This type of analysis has been performed for the LacI/PurR family bacterial transcription factors (Mirny and Gelfand 2002) and in two different families of eukaryotic transcription factors (basic leucine zippers and nuclear receptors) (Donald and Shakhnovich 2005). Comparison of more divergent family members can reveal essential elements of the protein's structural core.

1.2.2 Identification of homologous relationships

Homologous proteins have evolved from a common ancestor. Because homologs often perform similar functions and adopt similar structural folds, homologous relationships are arguably the most informative attribute of protein comparison in terms of understanding function, structure, and molecular evolution. Homology is most often established based on sequence similarity, although functional relatedness, shared catalytic residues or substrate-binding motifs, overall structural similarity, and the presence of distinct structural features are commonly used as verification. The identification of homology relationships is more often associated with the construction than with the utilization of classification schemes, although protein classification can potentially reveal remote homology links between highly divergent family members that might otherwise be found only by transitivity. Additionally, studying sets of homologs that are united in a protein classification scheme can help to establish criteria for distinguishing false positives from the true homologs of a particular protein family.

1.2.3 Functional inference

Another potential application of protein classification would be the inference of functional information about a newly discovered protein following the identification of homologs. Based on the observation that evolutionary relatives commonly have similar or related functions, possible biochemical roles for an uncharacterized protein can be inferred. This information could then be utilized to aid the experimental determination of a protein's role at the biochemical and cellular level.

Additionally, residues responsible for particular aspects of a protein's function can be predicted based on comparison of homologous sequences. Substrate binding, metal ion coordination, and catalytic roles are often accomplished by highly conserved charged or polar amino acids, which can often be identified by studying multiple alignments for a protein family.

1.2.4 Structure prediction

Based on the observation that homologous proteins generally adopt similar structural folds, classification schemes can be used to generate structural models for uncharacterized proteins. Potential uses might include producing models of active sites, substrate-binding sites, or other protein-protein interaction surfaces. These models can contribute to the prediction of functional residues, which can subsequently be tested experimentally. Additionally, structure models of enzyme active sites are commonly used in drug design.

1.3 DESCRIPTION OF DISSERTATION WORK

In this dissertation, protein classification is applied to the study of structural, functional, and evolutionary relationships within and between protein families. The work presented here addresses two general objectives. First, new classification schemes are constructed to gain a more complete understanding of sequence-structure-function relationships within large classes of proteins. Second, an algorithm is developed to make further use of valuable classification databases that already exist.

Three distinct protein classification projects are presented. These three projects are each based on the same two-tier protein classification hierarchy. The first level of this hierarchy reflects structural similarity among protein domains. Domains grouped at this level adopt structural folds that share a common architecture and topology. Consequently, this first tier is designated as the "fold group" level. The second level of the hierarchy signifies an evolutionary relationship (i.e. homology) between members and, in general, corresponds to sequence similarity. This second tier is referred to as the "family" level. Thus, proteins classified by this approach are evaluated in terms of both their structural and evolutionary relatedness.

In the first project, classification is applied to the study of a large functional class of proteins. Kinases are a large group of enzymes that catalyze the transfer of the terminal phosphate group from ATP to a small molecule, lipid, or protein substrate. Despite that all kinases catalyze essentially the same biochemical reaction (differing only in their substrate specificity), families of these proteins are known to adopt several different structural folds. This observation, coupled with the availability of thousands of kinase sequences, hundreds of kinase structures, and a wealth of biochemical data, makes kinases an ideal group of proteins for sequence-structure-function analysis. All available kinase sequences and structures have been organized according to the 2-tier classification hierarchy discussed above. The resulting classification scheme is subsequently used to study various aspects of kinase biochemistry and evolution, such as investigation into how different structural folds carry out the same fundamental aspects of the kinase phosphotransfer reaction. This work is described in Chapter 2.

The second project addresses the classification of a large structural, rather than functional, class of proteins. Small, disulfide-rich protein domains have global folds that are stabilized primarily by the formation of disulfide bonds, and to a much lesser extent by secondary structure and hydrophobic interactions. These domains typically lack a large hydrophobic core and have secondary structure elements that are small and irregular. In order to understand the structural and functional diversity among available small disulfide-rich proteins, these domains have been classified into the described two-tier hierarchy such that the proteins are arranged according to both their structural and evolutionary relatedness. This classification scheme describes more distant similarities between disulfide-rich protein sequences and structures than have been previously acknowledged. The comprehensive classification of available small, disulfide-stabilized protein structures is discussed in Chapter 3.

Although newly constructed classification schemes such as those presented for the kinases and disulfide-rich proteins can give further insight into the evolutionary, structural, and functional relatedness of proteins, there are many existing classification databases whose usefulness has by no means been exhausted. One example is the SCOP

database, which can provide valuable information about the evolutionary and structural neighbors of a query protein, but remains perpetually outdated. Chapter 4 discusses an algorithm developed to assign new protein queries to existing classification schemes and thereby extend the utility of databases such as SCOP. This algorithm demonstrates that automated sequence and structure comparison tools can be used to largely reproduce assignments to manually curated classification databases. Methods for automatic updates to existing classification schemes become increasingly important with the rapid growth in sequence and structure databases.

CHAPTER 2:
Classification of Kinase Sequences and Structures

2.1 INTRODUCTION

2.1.1 Background

Kinases are a ubiquitous group of enzymes that participate in a variety of cellular pathways. By definition, the common name kinase is applied to enzymes that catalyze the transfer of the terminal phosphate group from ATP to an acceptor, which can be a small molecule, lipid, or protein substrate. The cellular and physiological roles of kinases are diverse. Many kinases participate in signal transduction pathways, in which these enzymes are essential components. Other kinases are centrally involved in the metabolism of carbohydrates, lipids, nucleotides, amino acids, vitamins, and cofactors. Additionally, some kinases have roles in various other processes such as gene regulation, muscle contraction, and antibiotic resistance. Because of their universal roles in cellular processes, kinases are among the best-studied enzymes at the structural, biochemical, and cellular level. Despite that all kinases use the same phosphate donor (in most cases, ATP) and catalyze essentially the same phosphoryl transfer reaction, they display remarkable diversity in their structural folds and substrate recognition mechanisms. This is likely due to the extraordinarily diverse nature of the structures and properties of their substrates.

Evolutionary relationships are often identified by sequence analysis. However, even sequence similarity searches with powerful profile-based tools such as PSI-BLAST (Altschul, Madden et al. 1997) and HMMER (Eddy 1998) tend to miss more distant homologs. In some cases, comparative analysis of protein structural folds also allows for the inference of biochemical and biological functional properties. Structure analysis methods are able to detect evolutionary relationships that sequence similarity searches miss because protein structure conservation persists after sequence similarity disappears.

However, similarity of fold alone does not necessarily indicate a common ancestor. Furthermore, structural information is much less readily available than sequence information. The most effective route to the identification of homologs and the prediction of protein function is provided by the integration of sequence and structure data.

Currently, several protein classification schemes such as SCOP (Murzin, Brenner et al. 1995), CATH (Orengo, Michie et al. 1997), and Pfam (Bateman, Birney et al. 2000) have been developed for the purpose of cataloging all protein sequences and structures. This work presents the classification of a single group of proteins that catalyze a similar phosphoryl transfer reaction, and the subsequent examination of the relationships between the fold and biochemistry within this group. Such protein classification is in demand by the biologists because it is a useful tool for analyzing various aspects of sequence-structure-function relationships in proteins, such as structure prediction or identification of functionally important residues. The availability of thousands of kinase sequences and hundreds of kinase structures coupled with a wealth of biochemical data make kinases an ideal group of enzymes for such structural/functional classification and analysis.

2.1.2 Objectives

In order to investigate the relationship between structural fold and functional specificities in kinases, a comprehensive analysis of available kinase structures and sequences has been carried out. All kinase sequences have been classified into a two-tier hierarchy of fold groups and families. A number of hypothetical proteins in the database which may possess kinase activity were also predicted, and a large-scale structural prediction for kinases with unknown structures was performed.

Given the rapid increase in the sizes of sequence and structure databases, the utility of a protein classification scheme is directly dependant upon its propensity to remain stable over time. Ideally, the backbone of a classification scheme should not

require fundamental revisions with the inclusion of additional information. Therefore, three years after the original kinase survey was completed, an updated version was carried out. Also, fold predictions were performed for those kinase families currently lacking a homolog with solved structure. This update serves two important purposes: to validate the robustness of the initial kinase classification scheme and to present, for the first time, a complete structural annotation of this large functional class of proteins. Despite that the total number of available kinase sequences increased more than 3-fold (>59,000 in the updated survey), the framework of the original classification remains sufficient for describing all available kinase sequences.

The common structural features of each fold group of kinases and the families therein are described, emphasizing the shared catalysis and substrate binding mechanisms as well as variations within the same fold groups. In particular, this work attempts to address the questions of how different kinase structural folds accomplish the same required steps in the common phosphoryl transfer reaction and in some cases even recognize exactly the same substrate, and conversely, how kinases of the same fold recognize substrates with very different structures.

2.2 METHODS

2.2.1 Initial groupings of kinase sequences

A list of all Pfam profiles (Bateman, Birney et al. 2000) from version 5.4 and COGs from version 2 (Tatusov, Galperin et al. 2000) that describe catalytic kinase domains was constructed. For cases in which a COG's contents were completely contained within a Pfam profile, the COG was removed from the list to avoid redundancy. The reduced list contained 44 Pfam profiles and 12 COG sequence sets. One COG from version 3 (Tatusov, Natale et al. 2001) was later added, forming a total of 57 kinase profiles (44 from Pfam and 13 from COG). The hmmbuild and hmmcalibrate

programs of the HMMER2 package (Eddy 1998) were used to construct profiles for the 13 COG sequence sets. The hmmsearch program of the HMMER2 package was then used to assign sequences from the non-redundant (nr) protein database at NCBI (June 2001) to the kinase profiles (E-value cutoff 0.1).

Additionally, the GREFD program of the SEALS package (Walker and Koonin 1997) was used to extract all sequences from the nr (June 2001) for which the definition line contained the pattern "kinase". Three iterations of PSI-BLAST were run for each "kinase" sequence that was not already assigned to a profile. Any of these sequences that produced hits (E-value cutoff 0.001) to already-assigned sequences were subsequently placed in those profiles. The remaining unassigned "kinase" sequences (sequences with the word "kinase" in the definition line which were not placed in the Pfam/COG kinase profiles) were then manually filtered to remove fragments, non-kinase entries (e.g. kinase inhibitors), and non-catalytic entries (e.g. regulatory subunits). Such sequences were identified by their annotations in the nr and by their lengths being too short to cover the complete protein. In the case of non-kinase or non-catalytic entries, lack of kinase activity was confirmed based on either literature available concerning the sequences in question or on obvious homology to a protein with known non-kinase function.

Thus, after manual filtering any remaining entries are considered "true" catalytic kinase sequences that cannot be assigned to existing kinase profiles by automatic methods with the criteria described above. These remaining sequences were clustered by sequence similarity using the GROUPER program (score cutoff 50, single linkage) of the SEALS package. Multiple sequence alignments and secondary structure predictions were performed on these groups. Further sequence similarity searches with PSI-BLAST were carried out with somewhat relaxed thresholds and their results were inspected manually. Some of these initially unassigned groupings could then be merged into existing profiles (based on the presence of conserved catalytic residues, matching secondary structure predictions, and other distinguishing motifs), while others were placed into novel groupings.

2.2.2 Establishing families of homologous kinases

After all the kinase sequences had been assigned to Pfam/COG profiles or to novel groupings, PSI-BLAST was used to detect possible evolutionary links between these Pfam/COG profiles. Sequences from different Pfam/COG profiles with statistically significant similarities were identified and assembled into families. In other words, sequences from each Pfam/COG profile in the same family usually produce significant PSI-BLAST hits to each other. Therefore, homology is inferred to all sequences in the same family. In most cases, finding these links was trivial. In this study, a trivial link is defined as one that is established by 3 iterations of PSI-BLAST with E-value cutoff 0.001. For orphan Pfam/COG profiles or sequence groupings, multiple alignments were constructed in order to reveal conserved active site motifs. Secondary structure predictions with Jpred (Cuff, Clamp et al. 1998) and manual inspection of PSI-BLAST search results were also performed. In some cases, these orphan groupings can be placed into existing families with confidence. Others were assigned as novel kinase families.

For the purposes of this study, only those enzymes in EC2.7.1.- (phosphotransferases with an alcohol group as acceptor), EC2.7.2- (phosphotransferases with a carboxyl group as acceptor), EC2.7.3.- (phosphotransferases with a nitrogenous group as acceptor), and EC2.7.4.- (phosphotransferases with a phosphate group as acceptor) of the EC (Enzyme Commission) system are examined. Once the groupings had been made based on the profile and sequence similarity searches, a handful of other activities that are not a part of this range also fell neatly into the pre-existing groups. These enzymes were also added to this analysis. These added activities include EC4.1.1.32 (phosphoenolpyruvate carboxykinase - GTP), EC4.1.1.49 (phosphoenolpyruvate carboxykinase - ATP), EC2.7.9.1 (pyruvate, phosphate dikinase), EC2.7.9.2 (pyruvate, water dikinase), EC2.7.9.3 (selenide, water dikinase), EC2.7.6.2 (thiamin pyrophosphokinase), and EC2.7.6.3 (7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase). It should also be noted that many of the activities between EC2.7.1.- and EC2.7.4.- do not yet have identified sequences and therefore could not be

included in the groupings. Kinase activities utilizing different phosphate donors, such as EC2.7.2.10 (phosphoglycerate kinase (GTP)), were included in this classification if they were found to belong to a pre-existing kinase group. Kinases that do not hydrolyze ATP and do not belong to existing kinase groups were intentionally excluded. These excluded kinase activities include EC2.7.1.69 (protein-N(PI)-phosphohistidine-sugar phosphotransferase) and EC2.7.3.9 (phosphoenolpyruvate--protein phosphatase).

2.2.3 Fold group classification

In the original kinase survey, a total of 30 kinase families, containing as few as 2 sequences and up to as many as 9,600 sequences, were formed. 19 of these kinase families contained at least one member with a solved structure. These 19 families were assembled into 7 fold groups based on similarities of structural fold. Families in the same fold group share structurally similar nucleotide-binding domains that are of the same architecture and topology (or related by circular permutation) for at least the core of the domain. The remaining 10 groups were composed of the 11 families which, at that time, contained no members with solved structures.

2.2.4 Distribution of kinase sequences in the completely sequenced genomes

The distribution of kinase sequences in the completely sequenced genomes of several representative species was also investigated in the initial kinase survey. For each representative genome, the number of kinase sequences in each of the families was determined. The hmmsearch program (HMMER2) was used to assign sequences from the selected genomes to the 57 kinase profiles (E-value cutoff 0.1). In each family, any assigned sequences for which the word “kinase” was not found in the definition line were identified. BLAST was run for each of these potential non-kinase sequences in order to identify homologs. Any sequences for which BLAST results did not indicate that the protein was a kinase were removed from the profiles. Additionally, the GREFD program

(SEALS) was used to extract all entries with the word "kinase" in the definition line, and PSI-BLAST was used to place any unassigned "kinase" sequences into the profiles. The genomes of *Homo sapiens* (downloaded from ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/protein/, 12.13.2001), *Saccharomyces cerevisiae* (NC_001133, NC_001134, NC_001135, NC_001136, NC_001137, NC_001138, NC_001139, NC_001140, NC_001141, NC_001142, NC_001143, NC_001144, NC_001145, NC_001146, NC_001147, and NC_001148), *Escherichia coli* (NC_000913), and *Methanococcus jannaschii* (NC_000909) were obtained from the NCBI site. The *Drosophila melanogaster* genome (release 2) was downloaded from the Berkeley Drosophila Genome Project site (http://www.fruitfly.org/). The *Caenorhabditis elegans* genome (wormpep70) was downloaded from the Sanger Institute site (http://www.sanger.ac.uk/Projects/C_elegans/wormpep/).

2.2.5 Constructing updated families and fold groups of kinases

The updated version of the kinase classification scheme was assembled by the same strategy that was applied in the construction of the first kinase survey, with the previous classification used as a framework for this update. Briefly, the hmmsearch program of the HMMER2 package was used to assign sequences from the NCBI non-redundant (nr) database (July 2004) to the set of 57 profiles describing catalytic kinase domains (E-value cutoff 0.1). As these profiles had been assembled into families of homologous sequences in the initial classification scheme, the sequences assigned to these profiles by hmmsearch were then placed in the appropriate kinase families. The GREFD program of the SEALS package was again used to extract from the nr (July 2004) all sequences for which the definition line contained the pattern "kinase". PSI-BLAST hits were used to assign these kinases to families as described in section 2.2.1. Any remaining unassigned sequences were manually examined as described in section 2.2.1 and either removed or placed into existing families as appropriate. The lists of

newly identified kinase sequences were appended to the each of the kinase families included in the initial classification.

In the initial classification, fold groups were assembled based solely on similarity of structures. Families in the same fold group share structurally similar nucleotide-binding domains that are of the same architecture and topology (or related by circular permutation) for at least the core of the domain. Some of the recently solved kinase structures allowed for the merging of certain kinase families to previously established fold groups based on these same structural similarity guidelines.

The meaning of families and fold groups in the new version of the classification remains unaltered: the families contain homologous kinase sequences, while the fold groups imply similarity of structural fold but not homology.

2.2.6 Fold predictions

To provide fold assignments for the remaining structurally uncharacterized kinase families, initial analysis was performed with standard sequence similarity search methods such as transitive PSI-BLAST, RPS-BLAST, and profile HMMs from SMART (Letunic, Copley et al. 2004). All searches were initiated with the representative sequences of the families (Table 2.4). Transitive PSI-BLAST (E-value threshold 0.01) was run against the nr until convergence. The CDD (RPS-BLAST) and SMART (profile HMMs) web tools were used with default settings to detect distant homology to other conserved protein domains annotated in the SMART, Pfam, and COG databases. In addition, RPS-BLAST was exploited to compare query sequences directly to the PDB using the Gene Relational Database system (GRDB; http://basic.bioinfo.pl). Further analysis was carried out using Meta Server (Bujnicki, Elofsson et al. 2001), which assembles the results of various secondary structure prediction and fold recognition methods. Collected predictions were screened with 3D-Jury (Ginalski, Elofsson et al. 2003), the consensus method of fold recognition servers. The default servers used by the 3D-Jury system for consensus building include: ORFeus (Ginalski, Pas et al. 2003), Meta-BASIC (Ginalski, von

Grotthuss et al. 2004), FFAS03 (Rychlewski, Jaroszewski et al. 2000), mGenTHREADER (Jones 1999), INBGU (Fischer 2000), RAPTOR (Xu, Li et al. 2003), FUGUE-2 (Shi, Blundell et al. 2001), and 3D-PSSM (Kelley, MacCallum et al. 2000). Final fold/template selections were based on 3D-Jury reliability scores as well as those of individual servers, correctness of mapping of predicted and observed secondary structure elements, and conservation of functionally and/or structurally important residues. In the case of inositol 1,3,4,5,6-pentakisphosphate 2-kinase, initial fold assignment was based on functional analogy to 1-phosphatidylinositol-4-phoshate 5-kinase, which phosphorylates similar substrates.

Multiple sequence alignments for considered protein families were prepared using PCMA (Pei, Sadreyev et al. 2003) followed by manual adjustment. Sequence-to-structure alignments between analyzed kinase families and their distantly related template families were built using a consensus alignment approach and 3D assessment (Ginalski and Rychlewski 2003) based mainly on 3D-Jury results for representative kinase sequences. Sequences of distantly related proteins of known structure were aligned first based on the superposition of their 3D structures. In the case of inositol 1,3,4,5,6-pentakisphosphate 2-kinase, sequence-to-structure alignment was prepared manually with respect to the results of secondary structure predictions and the preservation of functionally critical residues as well as the hydrophobic core of the protein.

2.2.7 Alterations within the kinase classification

Although the framework of the classification remains essentially unchanged, the organization within the classification has been slightly modified. More specifically, the numbering of the fold groups has been adjusted so that all kinase families with unsolved structures are at the end. Furthermore, the EC numbers were updated to reflect the organization of the EC database in July 2004. Therefore, the EC content of each family may differ somewhat between the initial and updated classifications, but these changes do not indicate new additions to the family unless otherwise indicated.

2.3 RESULTS OF THE KINASE CLASSIFICATION

2.3.1 Results of the initial kinase classification (June 2001)

The initial kinase classification is summarized in Tables 2.1-2.4. For each fold group and family therein, the Pfam/COG members are listed as well as the kinase activities and a corresponding representative PDB or gi accession number. The total number of sequences in each family and group is specified as well. The EC activities in bold had a representative with solved structure in June 2001. It should be noted that the activity lists are not exhaustive, as they include only kinase activities with designated EC numbers (as of June 2001 for Tables 2.1-2.4). Of the 184 enzymes listed in the EC system over the chosen range (EC2.7.1.- through EC2.7.4.-) at that time, 112 activities were placed in the kinase families. Sequences for 70 of the remaining kinase activities from the chosen EC range were not identified at the time of the initial survey and thus could not be included in this analysis. The two remaining kinase activities were intentionally excluded (see section 2.2.2). Grey shading indicates proteins that were shown to be non-kinases when updating the kinase study.

Overall, 17,310 sequences were analyzed and classified into 30 families in the initial kinase survey. Sequences in each family are supported by statistically significant sequence similarities, indicating that they are homologs. Some of these families unify several Pfam/COG members. In the case of the P-loop kinase family, for example, 18 Pfam/COG members are found to contain statistically significant links and therefore belong to the same protein family. There were also 9 families, each containing between 2 and 148 sequences, which were not present in the versions of Pfam or COG used for the HMM searches of known catalytic kinase domains.

Families are assembled into fold groups based on similarity of structural fold. Within a fold group, the core of the nucleotide-binding domain of each family has the

same architecture, and the topology is either identical or related by circular permutation. Homology between families in a fold group is not implied. The structural features of the nucleotide-binding domain of each group are included in the fold group descriptions in section 2.4. Most of these kinase sequences (~98%) were found to belong to families with a structure representative known at the time of the initial kinase survey, and were placed in one of seven fold groups. The seven kinase fold groups included in the initial classification were all from either the α+β or the α/β class in SCOP, with approximately half of the families in these seven groups belonging to each of these two classes.

As previously mentioned, not all kinase activities specified in the EC are currently associated with an annotated sequence. Because the entire genomes of many model organisms have been sequenced, it is probable that the sequences for most, if not all genes encoding the remaining kinase activities are already known, but remain to be annotated or experimentally characterized. It is likely that most of these kinases would fall into one of the existing families or fold groups of the current classification scheme.

Table 2.1: Initial Kinase Classification, Fold Group 1

Fold Group	*Family and PFAM/COG members*	*Kinase Activities (E.C.)*	*Representative PDB*
Group 1: protein S/T-Y kinase/ atypical protein kinase/ lipid kinase/ ATP-grasp *9799 sequences*	protein S/T-Y kinase/ atypical protein kinase: COG0478, COG2112, PF00069, PF00454, PF01163, PF01633 *9600 sequences*	2.7.1.32 Choline kinase **2.7.1.37 Protein kinase** **2.7.1.38 Phosphorylase kinase** 2.7.1.39 Homoserine kinase 2.7.1.67 1-phosphatidylinositol 4-kinase 2.7.1.70 Protamine kinase 2.7.1.72 Streptomycin 6-kinase 2.7.1.82 Ethanolamine kinase 2.7.1.87 Streptomycin 3"-kinase **2.7.1.95 Kanamycin kinase** 2.7.1.100 5-methylthioribose kinase 2.7.1.103 Viomycin kinase **2.7.1.112 Protein-tyrosine kinase** 2.7.1.116 [Isocitrate dehydrogenase (NADP+)] kinase **2.7.1.117 [Myosin light-chain] kinase** 2.7.1.119 Hygromycin-B kinase **2.7.1.123 Calcium/calmodulin-dependent protein kinase** 2.7.1.125 Rhodopsin kinase 2.7.1.126 [Beta-adrenergic-receptor] kinase 2.7.1.129 [Myosin heavy-chain] kinase 2.7.1.135 [Tau protein] kinase **2.7.1.137 1-phosphatidylinositol 3-kinase** 2.7.1.141 [RNA-polymerase]-subunit kinase	PDB: 1cdk
	lipid kinase: PF01504 *82 sequences*	**2.7.1.68 1-phosphatidylinositol-4-phosphate kinase**	PDB: 1bo1
	ATP-grasp: PF01326 *117 sequences*	2.7.1.133 1D-myo-inositol-trisphosphate 6-kinase 2.7.1.139 1D-myo-inositol-trisphosphate 5-kinase **2.7.9.1 Pyruvate, phosphate dikinase** 2.7.9.2 Pyruvate, water dikinase	PDB: 1dik

Table 2.2: Initial Kinase Classification, Fold Group 2

Fold Group	*Family and PFAM/COG members*	*Kinase Activities (E.C.)*	*Representative PDB*
Group 2: Rossmann-like *3777 sequences*	P-loop kinases: COG0645, COG1618, COG1663, COG1936, COG2019, PF00265, PF00406, PF00485, PF00625, PF00693, PF01121, PF01202, PF01583, PF01591, PF01712, PF02223, PF02224, PF02283 *1756 sequences*	2.7.1.12 Gluconokinase **2.7.1.19 Phosphoribulokinase** **2.7.1.21 Thymidine kinase** **2.7.1.22 Ribosylnicotinamide kinase** **2.7.1.24 Dephospho-CoA kinase** **2.7.1.25 Adenylylsulfate kinase** **2.7.1.33 Pantothenate kinase** 2.7.1.37 Protein kinase (bacterial) 2.7.1.48 Uridine kinase **2.7.1.71 Shikimate kinase** 2.7.1.74 Deoxycytidine kinase 2.7.1.76 Deoxyadenosine kinase 2.7.1.78 Polynucleotide 5'-hydroxyl-kinase **2.7.1.105 6-phosphofructo-2-kinase** **2.7.1.113 Deoxyguanosine kinase** 2.7.1.130 Tetraacyldisaccharide 4'-kinase 2.7.4.2 Phosphomevalonate kinase **2.7.4.3 Adenylate kinase** 2.7.4.4 Nucleoside-phosphate kinase **2.7.4.8 Guanylate kinase** **2.7.4.9 Thymidylate kinase** **2.7.4.10 Nucleoside-triphosphate--adenylate kinase** **2.7.4.13 (Deoxy)nucleoside-phosphate kinase** **2.7.4.14 Cytidylate kinase** **2.7.4.- Uridylate kinase**	PDB: 1qf9
	phosphoenolpyruvate carboxykinase: COG1493, PF01293, PF00821 *212 sequences*	**2.7.1.37 Protein kinase (HPr kinase/phosphatase)** 4.1.1.32 Phosphoenolpyruvate carboxykinase (GTP) **4.1.1.49 Phosphoenolpyruvate carboxykinase (ATP)**	PDB: 1aq2
	phosphoglycerate kinase: PF00162 *271 sequences*	**2.7.2.3 Phosphoglycerate kinase** 2.7.2.10 Phosphoglycerate kinase (GTP)	PDB: 13pk
	aspartokinase: PF00696 *420 sequences*	**2.7.2.2 Carbamate kinase** 2.7.2.4 Aspartate kinase 2.7.2.8 Acetylglutamate kinase 2.7.2.11 Glutamate 5-kinase 2.7.4.- Uridylate kinase	PDB: 1b7b
	phosphofructokinase-like: PF00365, PF00781, PF01219, PF01513 *451 sequences*	**2.7.1.11 6-phosphofructokinase** 2.7.1.23 NAD(+) kinase 2.7.1.56 1-phosphofructokinase 2.7.1.90 Diphosphate--fructose-6-phosphate 1-phosphotransferase 2.7.1.107 Diacylglycerol kinase	PDB: 4pfk
	ribokinase-like: PF00294, PF01256, PF02110 *517 sequences*	**2.7.1.2 Glucokinase** 2.7.1.3 Ketohexokinase 2.7.1.4 Fructokinase 2.7.1.11 6-phosphofructokinase **2.7.1.15 Ribokinase** **2.7.1.20 Adenosine kinase** 2.7.1.35 Pyridoxal kinase 2.7.1.45 2-dehydro-3-deoxygluconokinase 2.7.1.49 Hydroxymethylpyrimidine kinase **2.7.1.50 Hydroxyethylthiazole kinase** 2.7.1.56 1-phosphofructokinase 2.7.1.73 Inosine kinase 2.7.1.144 Tagatose-6-phosphate kinase 2.7.1.146 ADP-dependent phosphofructokinase 2.7.1.147 ADP-dependent glucokinase 2.7.4.7 Phosphomethylpyrimidine kinase	PDB: 1rkd
	L-2-haloacid dehalogenase *2 sequences*	2.7.1.39 Homoserine kinase	PDB: 1j97
	thiamin pyrophosphokinase *148 sequences*	**2.7.6.2 Thiamin pyrophosphokinase**	PDB: 1ig0

Table 2.3: Initial Kinase Classification, Fold Groups 3-7

Fold Group	*Family and PFAM/COG members*	*Kinase Activities (E.C.)*	*Representative PDB*
Group 3: ferredoxin-like fold kinases *1798 sequences*	nucleoside-diphosphate kinase: PF00334 *200 sequences*	**2.7.4.6 Nucleoside-diphosphate kinase**	PDB: 2bef
	HPPK: PF01288 *70 sequences*	**2.7.6.3 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase**	PDB: 1eqo
	guanido kinases: PF00217 *151 sequences*	2.7.3.1 Guanidoacetate kinase **2.7.3.2 Creatine kinase** **2.7.3.3 Arginine kinase** 2.7.3.5 Lombricine kinase	PDB: 1bg0
	histidine kinase: PF00512, COG2172 *1377 sequences*	**2.7.1.37 Protein kinase (Histidine kinase)** **2.7.1.99 [Pyruvate dehydrogenase(lipoamide)] kinase** **2.7.1.115 [3-methyl-2-oxobutanoate dehydrogenase lipoamide)] kinase**	PDB: 1i59
Group 4: ribonuclease H-like *723 sequences*	COG0837, PF00349, PF00370, PF00871	**2.7.1.1 Hexokinase** 2.7.1.2 Glucokinase 2.7.1.4 Fructokinase 2.7.1.5 Rhamnulokinase 2.7.1.12 Gluconokinase 2.7.1.16 L-ribulokinase 2.7.1.17 Xylulokinase 2.7.1.27 Erythritol kinase **2.7.1.30 Glycerol kinase** 2.7.1.47 D-ribulokinase 2.7.1.51 L-fuculokinase 2.7.1.53 L-xylulokinase 2.7.1.55 Allose kinase 2.7.1.58 2-dehydro-3-deoxygalactonokinase 2.7.1.59 N-acetylglucosamine kinase 2.7.1.60 N-acylmannosamine kinase 2.7.1.63 Polyphosphate-glucose phosphotransferase 2.7.1.85 Beta-glucoside kinase **2.7.2.1 Acetate kinase** 2.7.2.7 Butyrate kinase 2.7.2.14 Branched-chain-fatty-acid kinase	PDB: 1dgk
Group 5: TIM β/α-barrel kinase *231 sequences*	PF00224	**2.7.1.40 Pyruvate kinase**	PDB: 1a49
Group 6: GHMP kinase *382 sequences*	COG1685, COG1907, PF00288, PF01971	2.7.1.6 Galactokinase **2.7.1.36 Mevalonate kinase** **2.7.1.39 Homoserine kinase** 2.7.1.71 Shikimate kinase 2.7.4.2 Phosphomevalonate kinase 2.7.1.148 4-(cytidine 5'-diphospho)-2-*C*-methyl-D-erythritol kinase	PDB: 1h72
Group 7: AIR synthetase-like *251 sequences*	PF00586	2.7.4.16 Thiamine-phosphate kinase 2.7.9.3 Selenide, water dikinase	PDB: 1cli

Table 2.4: Initial Kinase Classification, Fold Groups 8-17

Fold Group	*Family and PFAM/COG members*	*Kinase Activities (E.C.)*	*Representative gi*
Group 8: integral membrane kinases *63 sequences*	dolichol kinase: PF01879 *24 sequences*	2.7.1.108 Dolichol kinase	gi|6323655 [349..513]
	undecaprenol kinase *39 sequences*	2.7.1.66 Undecaprenol kinase	gi|1705428
Group 9: polyphosphate kinase *63 sequences*	PF02503	2.7.4.1 Polyphosphate kinase	gi|7465499 [48..730]
Group 10: riboflavin kinase *69 sequences*	PF01687	2.7.1.26 Riboflavin kinase	gi|6320442
fold group 11: inositol 1,4,5-trisphosphate 3-kinase *57 sequences*		2.7.1.127 1D-myo-inositol-trisphosphate 3-kinase	gi|10176869
Group 12: inositol 1,3,4,5,6-pentakisphosphate 2-kinase *2 sequences*		--	gi|6320521
Group 13: tagatose 6-phosphate kinase *8 sequences*		2.7.1.101 Tagatose kinase	gi|1168382
Group 14: pantothenate kinase *6 sequences*		2.7.1.33 Pantothenate kinase	gi|4191500
Group 15: glycerate kinase *20 sequences*		2.7.1.31 Glycerate kinase	gi|2495546
Group 16: putative glycerate kinase *18 sequences*	COG2379	2.7.1.31 Glycerate kinase	gi|1907334
Group 17: dihydroxyacetone kinase *43 sequences*		2.7.1.29 Glycerone kinase	gi|7387627

2.3.2 Results of the updated kinase classification (July 2004)

The updated kinase classification is summarized in Tables 2.5-2.7. The EC activities in bold have solved structures. Red bold entries indicate that the first structure associated with that kinase activity was solved after the initial kinase classification was completed. Underlined entries are new kinase activities that were not included in the initial survey, although activities added due to reorganization or updating of the EC database are not underlined. It should again be noted that the activity lists are not exhaustive, as they include only those kinase activities that have been annotated so far. Of the 190 kinases listed in the EC system over the chosen range (EC2.7.1.- through

EC2.7.4.-) in July 2004, 126 activities were placed in kinase families. Like in the initial survey, enzymes that use phosphate donors other than ATP are intentionally excluded, except when such kinases belong to existing families. Sequences for the remaining kinase activities have not been identified and thus are not included in this analysis, although it is possible that some of the unannotated kinases may be among the sequences with only general kinase function annotation (e.g. "similar to such-and-such kinase").

Overall, 59,402 sequences are classified into 25 families of homologous kinases. These kinase families are further assembled into 12 fold groups based on similarity of structural fold. 22 of the 25 families belong to 10 fold groups for which the structural fold is known. One additional family (polyphosphate kinase) is now associated with a predicted structural fold and presently forms a distinct fold group. The two remaining families are both integral membrane kinases and comprise the final fold group.

Table 2.5: Updated Kinase Classification, Fold Group 1

Fold Group	*Family and PFAM/COG members*	*Kinase Activity (E.C.)*	*Representative PDB or gi*
Group 1: protein S/T-Y kinase/ atypical protein kinase/ lipid kinase/ ATP-grasp *23124 sequences*	protein S/T-Y kinase/ atypical protein kinase: COG0478, COG2112, PF00069, PF00454, PF01163, PF01633 *22074 sequences*	**2.7.1.32 Choline kinase** **2.7.1.37 Protein kinase** **2.7.1.38 Phosphorylase kinase** 2.7.1.39 Homoserine kinase 2.7.1.67 1-phosphatidylinositol 4-kinase 2.7.1.72 Streptomycin 6-kinase 2.7.1.82 Ethanolamine kinase 2.7.1.87 Streptomycin 3"-kinase **2.7.1.95 Kanamycin kinase** 2.7.1.100 5-methylthioribose kinase 2.7.1.103 Viomycin kinase 2.7.1.109 [Hydroxymethylglutaryl-CoA reductase ($NADPH_2$)] kinase **2.7.1.112 Protein-tyrosine kinase** 2.7.1.116 [Isocitrate dehydrogenase (NADP+)] kinase **2.7.1.117 [Myosin light-chain] kinase** 2.7.1.119 Hygromycin-B kinase **2.7.1.123 Calcium/calmodulin-dependent protein kinase** 2.7.1.125 Rhodopsin kinase **2.7.1.126 [Beta-adrenergic-receptor] kinase** 2.7.1.129 [Myosin heavy-chain] kinase **2.7.1.135 [Tau protein] kinase** 2.7.1.136 Macrolide 2'-kinase **2.7.1.137 1-phosphatidylinositol 3-kinase** 2.7.1.141 [RNA-polymerase]-subunit kinase 2.7.1.153 Phosphatidylinositol-4,5-bisphosphate 3-kinase 2.7.1.154 Phosphatidylinositol-4-phosphate 3-kinase	PDB: 1cdk
	lipid kinase: PF01504 *321 sequences*	**2.7.1.68 1-phosphatidylinositol-4-phosphate 5-kinase** **2.7.1.127 1D-myo-inositol-trisphosphate 3-kinase** 2.7.1.140 Inositol-tetrakisphosphate 5-kinase 2.7.1.149 1-phosphatidylinositol 5-phosphate 4-kinase 2.7.1.150 1-phosphatidylinositol 3-phosphate 5-kinase 2.7.1.151 Inositol-polyphosphate multikinase 2.7.4.21 Inositol-hexakisphosphate kinase	PDB: 1bo1
	ATP-grasp: PF01326 *729 sequences*	2.7.1.134 Inositol-tetrakisphosphate 1-kinase **2.7.9.1 Pyruvate, phosphate dikinase** 2.7.9.2 Pyruvate, water dikinase	PDB: 1dik

Table 2.6: Updated Kinase Classification, Fold Group 2

Fold Group	*Family and PFAM/COG members*	*Kinase Activity (E.C.)*	*Representative PDB*
Group 2: Rossmann-like *17071 sequences*	P-loop kinases: COG0645, COG1618, COG1663, COG1936, COG2019, PF00265, PF00406, PF00485, PF00625, PF00693, PF01121, PF01202, PF01583, PF01591, PF01712, PF02223, PF02224, PF02283 *7732 sequences*	**2.7.1.12 Gluconokinase** **2.7.1.19 Phosphoribulokinase** **2.7.1.21 Thymidine kinase** **2.7.1.22 Ribosylnicotinamide kinase** **2.7.1.24 Dephospho-CoA kinase** **2.7.1.25 Adenylylsulfate kinase** **2.7.1.33 Pantothenate kinase** 2.7.1.37 Protein kinase (bacterial) **2.7.1.48 Uridine kinase** **2.7.1.71 Shikimate kinase** **2.7.1.74 Deoxycytidine kinase** 2.7.1.76 Deoxyadenosine kinase **2.7.1.78 Polynucleotide 5'-hydroxyl-kinase** **2.7.1.105 6-phosphofructo-2-kinase** **2.7.1.113 Deoxyguanosine kinase** 2.7.1.130 Tetraacyldisaccharide 4'-kinase **2.7.1.145 Deoxynucleoside kinase** **2.7.1.156 Adenosylcobinamide kinase** 2.7.4.1 Polyphosphate kinase 2.7.4.2 Phosphomevalonate kinase **2.7.4.3 Adenylate kinase** 2.7.4.4 Nucleoside-phosphate kinase **2.7.4.8 Guanylate kinase** **2.7.4.9 Thymidylate kinase** **2.7.4.10 Nucleoside-triphosphate--adenylate kinase** **2.7.4.13 (Deoxy)nucleoside-phosphate kinase** **2.7.4.14 Cytidylate kinase** **2.7.4.- Uridylate kinase**	PDB: 1qf9
	phosphoenolpyruvate carboxykinase: COG1493, PF01293, PF00821 *815 sequences*	**2.7.1.37 Protein kinase (HPr kinase/phosphatase)** **4.1.1.32 Phosphoenolpyruvate carboxykinase (GTP)** **4.1.1.49 Phosphoenolpyruvate carboxykinase (ATP)**	PDB: 1aq2
	phosphoglycerate kinase: PF00162 *1351 sequences*	**2.7.2.3 Phosphoglycerate kinase** 2.7.2.10 Phosphoglycerate kinase (GTP)	PDB: 13pk
	aspartokinase: PF00696 *2171 sequences*	**2.7.2.2 Carbamate kinase** 2.7.2.4 Aspartate kinase **2.7.2.8 Acetylglutamate kinase** 2.7.2.11 Glutamate 5-kinase 2.7.4.- Uridylate kinase	PDB: 1b7b
	phosphofructokinase-like: PF00365, PF00781, PF01219, PF01513 *1998 sequences*	**2.7.1.11 6-phosphofructokinase** **2.7.1.23 NAD(+) kinase** 2.7.1.56 1-phosphofructokinase **2.7.1.90 Diphosphate--fructose-6-phosphate 1- phosphotransferase** 2.7.1.91 Sphinganine kinase 2.7.1.107 Diacylglycerol kinase 2.7.1.138 Ceramide kinase	PDB: 4pfk
	ribokinase-like: PF00294, PF01256, PF02110 *2722 sequences*	**2.7.1.2 Glucokinase** 2.7.1.3 Ketohexokinase 2.7.1.4 Fructokinase 2.7.1.11 6-phosphofructokinase **2.7.1.15 Ribokinase** **2.7.1.20 Adenosine kinase** **2.7.1.35 Pyridoxal kinase** 2.7.1.45 2-dehydro-3-deoxygluconokinase 2.7.1.49 Hydroxymethylpyrimidine kinase **2.7.1.50 Hydroxyethylthiazole kinase** 2.7.1.56 1-phosphofructokinase 2.7.1.73 Inosine kinase 2.7.1.92 5-dehydro-2-deoxygluconokinase 2.7.1.144 Tagatose-6-phosphate kinase 2.7.1.146 ADP-dependent phosphofructokinase **2.7.1.147 ADP-dependent glucokinase** **2.7.4.7 Phosphomethylpyrimidine kinase**	PDB: 1rkd
	thiamin pyrophosphokinase *175 sequences*	**2.7.6.2 Thiamin pyrophosphokinase**	PDB: 1ig0
	glycerate kinase (previously Group 15) *107 sequences*	**2.7.1.31 Glycerate kinase**	PDB: 1to6

Table 2.7: Updated Kinase Classification, Fold Groups 3-12

Fold Group	*Family and PFAM/COG members*	*Kinase Activity (E.C.)*	*Representative PDB or gi*
Group 3: ferredoxin-like fold kinases *10973 sequences*	nucleoside-diphosphate kinase: PF00334 *923 sequences*	**2.7.4.6 Nucleoside-diphosphate kinase**	PDB: 2bef
	HPPK: PF01288 *609 sequences*	**2.7.6.3 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase**	PDB: 1eqo
	guanido kinases: PF00217 *324 sequences*	2.7.3.1 Guanidoacetate kinase **2.7.3.2 Creatine kinase** **2.7.3.3 Arginine kinase** 2.7.3.5 Lombricine kinase	PDB: 1bg0
	histidine kinase: PF00512, COG2172 *9117 sequences*	**2.7.1.37 Protein kinase (Histidine kinase)** **2.7.1.99 [Pyruvate dehydrogenase(lipoamide)] kinase** **2.7.1.115 3-methyl-2-oxobutanoate dehydrogenase(lipoamide) kinase**	PDB: 1i59
Group 4: ribonuclease H-like *2768 sequences*	COG0837, PF00349, PF00370, PF00871	**2.7.1.1 Hexokinase** 2.7.1.2 Glucokinase 2.7.1.4 Fructokinase 2.7.1.5 Rhamnulokinase 2.7.1.7 Mannokinase 2.7.1.12 Gluconokinase 2.7.1.16 L-ribulokinase 2.7.1.17 Xylulokinase 2.7.1.27 Erythritol kinase **2.7.1.30 Glycerol kinase** 2.7.1.33 Pantothenate kinase 2.7.1.47 D-ribulokinase 2.7.1.51 L-fuculokinase 2.7.1.53 L-xylulokinase 2.7.1.55 Allose kinase 2.7.1.58 2-dehydro-3-deoxygalactonokinase 2.7.1.59 N-acetylglucosamine kinase 2.7.1.60 N-acylmannosamine kinase 2.7.1.63 Polyphosphate-glucose phosphotransferase 2.7.1.85 Beta-glucoside kinase **2.7.2.1 Acetate kinase** 2.7.2.7 Butyrate kinase 2.7.2.14 Branched-chain-fatty-acid kinase 2.7.2.- Propionate kinase	PDB: 1dgk
Group 5: TIM β/α-barrel kinase *1119 sequences*	PF00224	**2.7.1.40 Pyruvate kinase**	PDB: 1a49
Group 6: GHMP kinase *885 sequences*	COG1685, COG1907, PF00288, PF01971	**2.7.1.6 Galactokinase** **2.7.1.36 Mevalonate kinase** **2.7.1.39 Homoserine kinase** 2.7.1.46 L-arabinokinase 2.7.1.52 Fucokinase 2.7.1.71 Shikimate kinase **2.7.1.148 4-(cytidine 5'-diphospho)-2-*C*-methyl-D-erythritol kinase** **2.7.4.2 Phosphomevalonate kinase**	PDB: 1h72
Group 7: AIR synthetase-like *1843 sequences*	PF00586	2.7.4.16 Thiamine-phosphate kinase 2.7.9.3 Selenide, water dikinase	PDB: 1cli
Group 8: riboflavin kinase (previously Group 10) *565 sequences*	PF01687	**2.7.1.26 Riboflavin kinase**	PDB: 1nb9
Group 9: dihydroxyacetone kinase (previously Group 17) *197 sequences*		**2.7.1.29 Glycerone kinase**	PDB: 1un9
Group 10: putative glycerate kinase (previously Group 16) *148 sequences*	COG2379	**2.7.1.31 Glycerate kinase**	PDB: 1o0u
Group 11: polyphosphate kinase (previously Group 9) *446 sequences*	PF02503	2.7.4.1 Polyphosphate kinase	gi|7465499 [48..730]
Group 12: integral membrane kinase (previously Group 8) *263 sequences*	dolichol kinase: PF01879 *127 sequences*	2.7.1.108 Dolichol kinase	gi|6323655 [349..513]
	undecaprenol kinase *136 sequences*	2.7.1.66 Undecaprenol kinase	gi|1705428

2.3.3 Framework of the classification remains unchanged

The updated classification includes 42,092 additional sequences, 343 additional kinase structures, and 12 additional kinase specificities compared to the original classification. Although the total number of kinase sequences included in the classification has an impressive increase of more than 3-fold (from 17,310 to 59,402), all new kinase sequences were found to be homologous to the previously established families, and thus are contained in the existing family and fold group classification. Furthermore, 343 additional kinase structures were solved since the initial classification was completed. The majority of these structures correspond to kinases for which at least one representative structure was already known. For example, dozens of additional eukaryotic protein serine-threonine/tyrosine kinase structures were solved. Structures of 23 kinases with previously uncharacterized structures were also published (shown in red bold text in Tables 2.5-2.7). The structural folds for 18 of these 23 kinases were predicted by the initial kinase classification based on their homology to proteins with known structures. All 18 of these predicted folds were shown be to correct by the experimentally determined structures. For example, choline kinase was expected to have a protein kinase-like fold similar to the other members of Family 1a (protein S/T-Y kinases and atypical protein kinases). The crystal structure of choline kinase (Peisach, Gee et al. 2003) shows that this protein does indeed adopt a eukaryotic protein kinase-like fold. Likewise, pyridoxal kinase was shown to have a ribokinase-like fold (Li, Kwok et al. 2002), as was predicted in the initial kinase classification. Thus, the placements of these kinases in the classification scheme remain unchanged.

The five remaining kinases with recently solved structures belong to families for which the fold was previously unknown. Two of these kinases, riboflavin kinase and dihydroxyacetone kinase, represent two new unique kinase folds. One glycerate kinase family, which was previously listed as an independent fold group, is now placed as an additional family in the Rossmann-like fold group due to similarities in architecture and topology of the predicted nucleotide-binding domain. As the nucleotide-binding domain

cannot be confidently predicted for a second distinct glycerate kinase family, these sequences tentatively remain as a separate fold group. Lastly, inositol 1,4,5-trisphosphate 3-kinase is now known to be a member of the lipid kinase-like family (Family 1b).

The total numbers of sequences, structures, families, and fold groups in the initial and updated classifications are summarized in Table 2.8.

Table 2.8: Comparison of Initial and Updated Kinase Surveys

	Sequences	***Structures***	***Families***	***Families with Known Structure***	***Fold Groups***	***Fold Groups with Known Structure***
Initial Survey	17310	359	30	19	17	7
Updated Survey	59402	702	25	22	12	10

2.3.4 Additional kinase activities in the updated classification

The updated classification includes 12 additional kinase activities. However, 7 of these activities reflect changes within the EC database rather than newly characterized kinase sequences. For example, while the structure of adenosylcobinamide kinase was published in 1998, its EC number (EC 2.7.1.156) was only assigned in April 2004. The sequences and structures of this kinase were already included in initial kinase survey (e.g. pdb|1cbu (Thompson, Thomas et al. 1998)) and were placed in the P-loop kinase family of the Rossmann-like fold group.

The updated kinase classification does include 5 newly annotated or characterized kinases (indicated by underlining in Tables 2.5-2.7). The first sequences associated with each of these activities (2.7.1.52 Fucokinase, 2.7.1.92 5-dehydro-2-deoxygluconokinase, 2.7.1.138 Ceramide kinase, and 2.7.1.140 Inositol-tetrakisphosphate 5-kinase) were identified after the initial kinase survey was completed. Sequences with [Hydroxymethylglutaryl-CoA reductase ($NADPH_2$)] kinase activity (2.7.1.109) were included in the initial classification, although only a general kinase activity ("AMP-

activated protein kinase") was assigned at the time. Thus, the specific kinase activity of this enzyme is a new addition to the kinase classification as well.

All of these newly annotated kinases belong to existing kinase families containing members that are well characterized both biochemically and structurally. The link between these kinases and members of the existing families can all be identified by BLAST with E-values less than $1e^{-5}$. Therefore, the catalytic mechanisms of these newly annotated kinases may be inferred from their closely related homologs.

2.4 DESCRIPTION OF KINASE FOLD GROUPS AND FAMILIES

2.4.1 Group 1: protein kinase-like

Group 1 is composed of three families: the protein serine/threonine-tyrosine kinase-like family (hereafter referred to as the protein kinase family), the lipid kinase family, and the ATP-grasp family. An evolutionary link between these three families based on structural similarity has been described previously (Grishin 1999). In each of the Group 1 families, the active site of the enzyme is located in the cleft between two α+β domains (Figure 2.1a). The protein kinase-like family members and lipid kinase family members have very similar N-terminal domains, but their C-terminal domains are different except for a region containing a 3-stranded antiparallel β-sheet and associated α-helices, which includes the αββ unit that is essential for nucleotide binding (Grishin 1999). The lipid kinase and ATP-grasp proteins, on the other hand, share very similar C-terminal domains but have different N-terminal domains. All three families bind ATP along the β-sheet of the αββ unit core of the C-terminal domain (Grishin 1999).

Family 1a: protein serine/threonine-tyrosine kinase-like

The protein kinase family has 6 Pfam/COG members. The majority of the sequences in this family are eukaryotic protein S/T-Y kinases. However, protein kinase

homologs in bacterial and archaeal species have also been identified (Leonard, Aravind et al. 1998). Links between the 6 Pfam/COGs are trivial. Two families have a trivial link if at least one sequence in one family produces a significant hit to at least one sequence in the other family within 3 iterations of PSI-BLAST (E-value cutoff 0.001). In this family, several sequences in each of the Pfam/COGs give significant PSI-BLAST hits to multiple sequences in PF00069 (protein kinase domain), which is the largest of the 6 Pfam/COG members in this family. In addition, there are several kinases with different specificities that can be linked to the protein kinase family but are not assigned by automatic methods. These sequences can be identified by the presence of conserved active site residues, which are well described for the protein kinase family (Goldsmith and Cobb 1994; Hanks and Hunter 1995), and by the predicted secondary structure patterns (Figure 2.1b). These proteins include hygromycin-B kinase, streptomycin 3"-kinase, fructosamine-3-kinase, isocitrate dehydrogenase kinase/phosphatase (Oudot, Cortay et al. 2001), kanamycin kinase, viomycin kinase, actin-fragmin kinase, homoserine kinase from several proteobacterial species, the kinase domain of ChaK (a transient receptor potential channel), eukaryotic elongation factor-2 kinase, and myosin heavy chain kinase. Solved structures for kanamycin kinase (Burk, Hon et al. 2001), actin-fragmin kinase (Steinbacher, Hof et al. 1999), and the kinase domain of ChaK (Yamaguchi, Matsushita et al. 2001) confirm that these three kinases are indeed structurally similar to protein kinases. The kinase domain of ChaK, eukaryotic elongation factor-2 kinase, and myosin heavy chain kinase are also known as α-kinases or atypical kinases. They have only limited sequence similarity to classical protein kinases (Ryazanov, Ward et al. 1997), so the structural similarity between α-kinases and classical protein kinases was unexpected.

Protein kinases are among the most thoroughly studied protein families. This family has several highly conserved active site residues. A conserved lysine residue from the N-terminal domain interacts with the α- and β-phosphates of ATP. The aspartate residue and the asparagine residue in a highly conserved DXXXXN motif play a role in catalysis and in coordinating a secondary Mg^{2+} cation, respectively. The primary Mg^{2+} cation is coordinated by the conserved aspartate residue of the DFG motif (Figure 2.1b).

Protein kinases also have a glycine-rich loop (GXGXXGXV) that interacts with the phosphates of the ATP.

The mechanism of protein kinases was historically thought to be acid-base catalysis. However, more recent studies have questioned this hypothesis and proposed that phosphoryl transfer is accomplished by the simultaneous transfer of a proton from the substrate hydroxyl group to an oxygen of the γ-phosphate (Hart, Hillier et al. 1998). The conserved aspartate residue (of the DXXXXN motif) that was once thought to act as the base catalyst is now suggested to have a role in stabilization of the protonated form of the transferred phosphate.

Figure 2.1: The Protein Kinase-Like Family

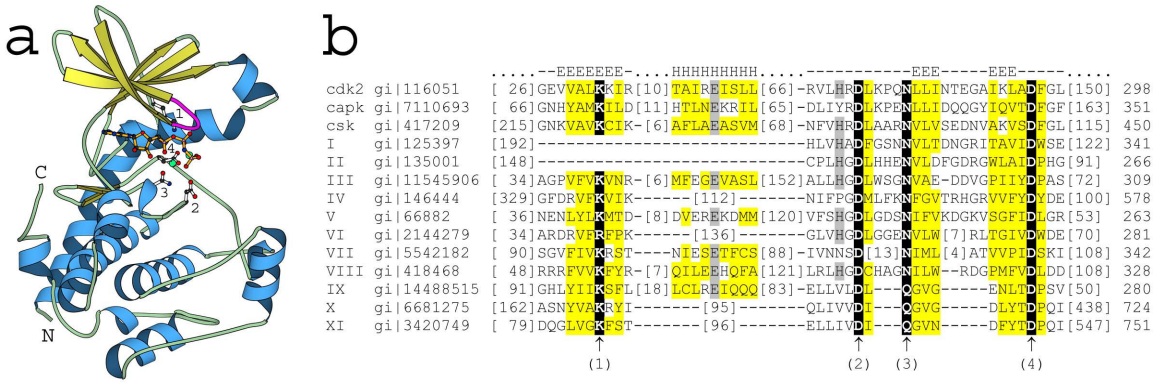


Figure 2.1: The Protein Kinase-Like and ATP-grasp Families. a) The protein kinase fold (cAMP dependent protein kinase, pdb|1cdk (Bossemeyer, Engh et al. 1993)). Residue 1 (Lys72) interacts with the α- and β- phosphates of ATP. Residue 2 (Asp166) is believed to have a role in catalysis. Residues 3 and 4 (Asn171 and Asp184, respectively) each coordinate a Mg^{2+} cation. The glycine-rich loop is shown in magenta. The nucleotide is orange and the Mg^{2+} cation is a green ball. All ribbon diagrams are made in MOLSCRIPT (Kraulis 1991). b) Addition of distant members to the protein kinase family. The first three sequences are members of the PK family with known structures: cell division protein kinase 2 (cdk2), cAMP dependent protein kinase (capk), tyrosine-protein kinase CSK (csk). I-IX are representative sequences of kinase groupings that were not assigned by automatic methods to any of the kinase profiles according to the criteria used for analysis, but are in fact evolutionarily related to the PK family: hygromycin-B kinase (I), streptomycin 3"-kinase (II), fructosamine-3-kinase (III), isocitrate dehydrogenase kinase/phosphatase (IV), kanamycin kinase (V), viomycin kinase (VI), actin-fragmin kinase (VII), homoserine kinase (VIII), the kinase domain of ChaK (IX), eukaryotic elongation factor-2 kinase (X), and myosin heavy chain kinase (XI). Conserved residues known to be involved in catalysis or substrate binding are highlighted black and shown in white bold letters; other highly conserved residues are highlighted grey. Uncharged residues in mainly hydrophobic sites are highlighted yellow. The numbers to the far right after each sequence indicate the total amino acid length of the sequence. Numbers in brackets specify the number of residues in an insertion that are not shown. Secondary structure is indicated above the alignment, with E signifying β-strands and H signifying α-helices.

Family 1b: lipid kinase

In the initial survey, the only identified kinase member of this family was type IIβ phosphatidylinositol phosphate kinase (PIPK). Since the N-terminal domains of PIPK and protein kinases are very similar, the glycine rich phosphate binding loop and the conserved lysine residue located at the N-terminal domain are preserved in both structures. Additionally, PIPK has two conserved aspartate residues, Asp278 and Asp369, which can be aligned with the catalytic aspartate of the DXXXXN motif, and the Mg^{2+} coordinating aspartate of the DFG motif in protein kinases, respectively (Rao, Misra et al. 1998; Grishin 1999). The roles of these residues are expected to be the same as those of their protein kinase counterparts.

In the updated survey, it was determined that inositol 1,4,5-trisphosphate 3-kinase (I3P3K; previously Group 11) and 1,3,4,5,6-pentakisphosphate 2-kinase (I5P2K; previously Group 12) are members of this family as well. I3P3K and I5P2K both catalyze phosphorylation reactions in the production of inositol polyphosphate (IP) second messengers. These kinases were placed in separate fold groups in the initial survey based on a lack of significant sequence similarity to each other or any other known kinase family. The solved structure of I3P3K and the predicted structure of I5P2K reveal similarity between these proteins and the lipid kinase family.

The solved structures of human I3P3K (pdb|1w2c (Gonzalez, Schell et al. 2004)) and rat I3P3K (pdb|1tzd (Miller and Hurley 2004)) reveal that the catalytic core of this kinase adopts a protein kinase-like fold and is comprised of two domains with the active site cleft in between (Figure 2.2a). The overall structure of I3P3K is most similar to the lipid kinases (pdb|1bo1 (Rao, Misra et al. 1998)) and also shares some similarity with the eukaryotic protein kinases of Family 1a. The shared structural core between lipid kinases and protein kinases includes all elements of the I3P3K N-terminal domain, and part of the β-sheet (β-strands 1, 4, 5) and the three α-helices of the C-terminal domain (Figure 2.2a).

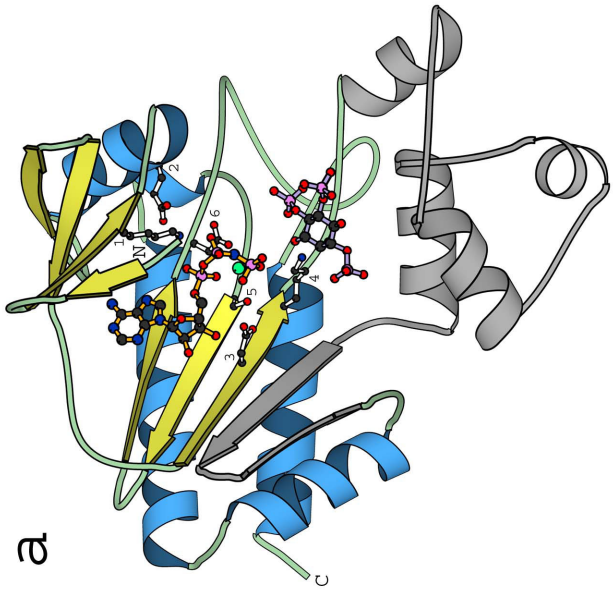
The mode of nucleotide binding in I3P3K is also very similar to that of eukaryotic protein kinases, as each of the critical nucleotide binding and Mg^{2+} binding residues in I3P3K plays the same role as a corresponding protein kinase residue. Lys209 (human

I3P3K or hI3P3K; residue **1** in Figure 2.2) forms a hydrogen bond with the α- and β-phosphates of ATP and corresponds to the highly conserved Lys72 in protein kinase A (PKA). This lysine residue is oriented by Glu215 in hI3P3K (residue **2** in Figure 2.2), corresponding to Glu91 in PKA. A second highly conserved lysine residue (Lys264 in hI3P3K; residue **4** in Figure 2.2) interacts with the 3-OH phosphate acceptor group of the inositol 1,4,5-trisphosphate substrate and likely stabilizes the γ-phosphate during transfer, similar to the role of Lys168 in PKA. Although Lys264 (hI3P3K) and Lys168 (PKA) are contributed by different structural elements in different regions of the protein sequence, these residues are found in equivalent spatial locations and likely play the same role in catalysis. A Mn^{2+} ion is coordinated by Asp416 (residue **6** is Figure 2.2), which corresponds to the conserved magnesium-binding residue Asp184 of the DFG motif in PKA. Ser399 (residue **5** in Figure 2.2) is expected to coordinate a second divalent cation that is not modeled in the I3P3K structure, as this residue is found in the equivalent spatial location as the conserved magnesium-binding residue Asn171 in PKA. These active site similarities also extend to other members of the lipid kinase family, such as phosphatidylinositol phosphate kinase IIβ (PIPK; pdb|1bo1), although a representative structure with bound nucleotide has not yet been solved.

Although I3P3K shares similarity of the overall fold as well as the active site with the related lipid kinase and protein kinase-like families, I3P3K is more closely related to the lipid kinase family. I3P3K and the lipid kinases share conserved motifs which are not found in protein kinases, including the substrate-binding/catalytic "DLK" motif (Asp262 to Lys264 in hI3P3K; Asp262 and Lys264 are residues **3** and **4** in Figure 2.2, respectively) and the magnesium-binding "SSLL" motif (Ser398 to Leu401 in hI3P3K; Ser399 is residue **5** in Figure 2.2). Additionally, Dali (Holm and Sander 1995) identifies lipid kinase representative PIPK (pdb|1bo1) as the closest structural neighbor of I3P3K (Miller and Hurley 2004). Thus, based on the similarity of structural fold and the conservation of critical nucleotide-binding, magnesium-binding, and catalytic residues, I3P3K can be assigned to the lipid kinase family (Family 1b) despite the lack of significant overall sequence similarity.

Figure 2.2: Structure of Inositol Polyphosphate Kinases

a) I3P3K (pdb|1w2c) adopts a lipid kinase-like fold. The common core of lipid kinases, eukaryotic protein kinases, and I3P3K is shown in color; additional elements are grey. Critical active site residues are 1 - Lys209, 2 - Glu215, 3 - Asp262, 4 - Lys264, 5 - Ser399, and 6 - Asp416. The nucleotide is orange, inositol 1,4,5-trisphosphate is purple, and the Mn^{2+} cation is a green ball. b) Alignment of I3P3K (gi|10176869, pdb|1w2c, pdb|1tzd) and I5P2K (gi|6320521) with two related kinase families: lipid kinase representative PIPK (pdb|1bo1; Family 1b) and protein kinase representative twitchin kinase (pdb|1koa; Family 1a). Italics denote α-helical regions for which the register of structure-based alignment cannot be obtained unequivocally due to significant structural divergence. Critical active site residues are in white bold text and highlighted black/magenta. Magenta highlighting indicates residues in equivalent spatial locations that perform equivalent roles, but do not align closely in sequence. Sequences are labeled by NCBI gi number or PDB code and an abbreviation of the species name. First and last residue numbers are given before and after each sequence. Numbers of excluded residues are specified in square brackets. Residue conservation is denoted by yellow highlighting (mostly hydrophobic positions), grey highlighting (mostly charged/polar positions), red bold text (small residues). Predicted (gi|6320521, gi|10176869) and observed (for PDB structures) secondary structure elements (E, β-strand; H, α-helix) are marked above the sequences (except gi|10176869 which is below the sequence) in italics and normal font, respectively. Abbreviations in this alignment are as follows: Ag *Anopheles gambiae*, Am *Apis mellifera*, At *Arabidopsis thaliana*, Ce *Caenorhabditis elegans*, Dh *Debaryomyces hansenii*, Dm *Drosophila melanogaster*, Gg *Gallus gallus*, Gz *Gibberella zeae*, Hs *Homo sapiens*, Mg *Mytilus galloprovincialis*, Mm *Mus musculus*, Rn *Rattus norvegicus*, Sc *Saccharomyces cerevisiae*.



Before the crystal structures of I3P3K were reported, fold predictions for both I3P3K and I5P2P were carried out. The results of fold predictions guided by 3D-Jury (Ginalski, Elofsson et al. 2003), secondary structure predictions, and observed presence of critical conserved sequence motifs indicated that both of these IP kinases would likely adopt a structural fold similar to lipid kinases and eukaryotic protein kinases, which are possibly related families that share a common ATP-binding site and structural core (Grishin 1999). Based on the structure predictions, a multiple alignment of representative I3P3K, I5P2K, lipid kinase, and protein kinase sequences was constructed (Figure 2.2b). This alignment shows that the critical functional residues (residues **1** - **6**) in these proteins are also conserved in the IP kinases. Furthermore, both I3P3K and I5P2K also have a predicted glycine-rich loop in the N-terminal region of the protein. From the multiple sequence alignment, it becomes apparent that I3P3K and I5P2K are more closely related to the other lipid kinases than to protein kinases. In addition to phosphorylating similar substrates, the IP kinases and lipid kinase family members each have critical active site lysine residue involved in stabilizing the γ-phosphate of ATP during transfer (residue **4** in Figure 2.2b) that has migrated in the sequence/structure relative to the protein kinase-like family. The solved structures of I3P3K from human (Gonzalez, Schell et al. 2004) and rat (Miller and Hurley 2004) confirm this non-trivial fold assignment as well as the predicted functional roles played by the conserved active site residues. Additionally, this further increases confidence in the I5P2K prediction. Thus, I3P3K and I5P2K are now assigned as members of the lipid kinase-like family (Family 1b) in the kinase classification.

Family 1c: ATP-grasp

The ATP-grasp fold describes several different ATP-hydrolyzing enzymes (Murzin 1996). However, there is only one kinase Pfam member of this family (PF01326: Pyruvate phosphate dikinase, PEP/pyruvate binding domain). In this family, ATP is held between two antiparallel β-sheets. Hence, the term 'ATP-grasp' is used to describe this fold in SCOP. The mechanism of the phosphotransfer reaction in pyruvate

phosphate dikinase involves the reversible phosphorylation of a histidine residue (Spronk, Yoshida et al. 1976; Goss, Evans et al. 1980). In this enzyme, the binding sites of the nucleotide and the pyruvate are in distant locations on the protein, and the shuttling of the phosphorylated histidine residue between the two active sites is accomplished by a dramatic swivel of the phospho-histidine domain (Herzberg, Chen et al. 1996). Inspection of the multiple alignment of inositol 1,3,4-trisphosphate 5/6-kinase sequences has indicated that this kinase also belongs to the ATP-grasp family. Figure 2.3 shows a sequence alignment of the large subunit of *Escherichia coli*, human, and yeast carbomoyl-phosphate synthases (a representative of the ATP-grasp fold) with inositol 1,3,4-trisphosphate 5/6-kinases.

Figure 2.3: Addition of Distant Members of the ATP-grasp Family

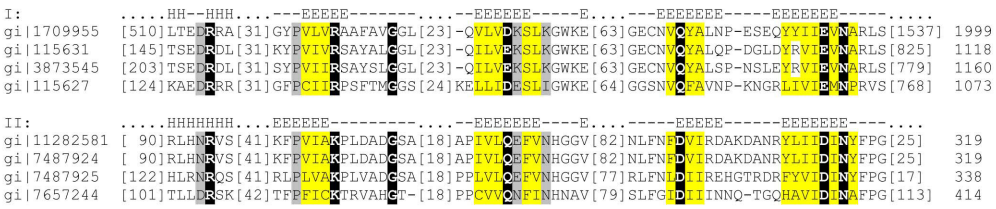


Figure 2.3: Addition of Distant Members of the ATP-grasp Family. I: representative sequences of the ATP-grasp fold (large subunit of carbamoyl-phosphate synthase); II: representative sequences of inositol 1,3,4-trisphosphate 5/6 kinase. Details pertaining to alignment layout are described in Figure 2.1 legend.

2.4.2 Group 2: Rossmann-like kinases

Group 2 includes 8 Rossmann-like kinase families. Eight families were identified in the initial survey. In the updated survey, one of these families was removed (HAD-like homoserine kinase/phosphoserine phosphatase) and another family was added (glycerate kinase, previously Group 15). The common structural feature of these families is that the architecture of their nucleotide-binding domain core is 3 layers (α/β/α) composed of βα repeats, with the central β-sheet mostly parallel. There is always a

change in direction of strand order in the middle of the β-sheet (Rossmann, Moras et al. 1974), resulting in the most common strand order of 321456, although modifications of this topology are frequent. The total number of β-strands and the strand order of the β-sheet differ between some families in this group. There is also a wide range of insertions or additional domains that are mostly associated with phosphoryl-acceptor substrate binding, accounting for the extremely diverse substrate specificities in this group of kinases. In addition to sharing a common fold of the nucleotide-binding domain, the families in the Rossmann-like group also have similar nucleotide-binding patterns. In each family, the nucleotide binds at the C-terminal end of the β-sheet, with the phosphate tail always located at the N-terminal end of one or more α-helices (Figures 2.4 and 2.5). More thorough descriptions of nucleotide-binding specifics for the families within this group are provided below.

Family 2a: P-loop kinases

The largest family in Group 2 is the P-loop containing kinases, which unifies 18 Pfam/COG members. 16 of these Pfam/COG members have trivial PSI-BLAST links. Alignments identifying the Walker-A and Walker-B motifs for the remaining two members are shown in Figure 2.4a. Additionally, a small number of phosphomevalonate kinase sequences from animals can be assigned to this family (Smit and Mushegian 2000). Alignments identifying the conserved diagnostic motifs for these phosphomevalonate kinase sequences are also shown in Figure 2.4a.

The P-loop kinases contain one 3-layered (α/β/α) domain. For the majority of the members of this family, the central parallel β-sheet is 5-stranded with strand order 23145. Nucleotide binding in this family is distinguished by the presence of the conserved Walker-A (GXXXXGKT/S) and Walker-B (ZZZZD, where Z is any hydrophobic residue) motifs (Walker, Saraste et al. 1982). The Walker-A motif forms a phosphate-binding loop (P-loop) and is found in a variety of different proteins that bind nucleotides (Saraste, Sibbald et al. 1990). In this family of kinases, the P-loop is located at the end of the first β-strand and includes the first half turn of the following α-helix. The conserved

lysine of the Walker-A motif binds to and orients oxygens of the β- and γ-phosphates of ATP. The essential Mg^{2+} cation is coordinated directly by the hydroxyl group of the conserved threonine/serine of the Walker-A motif and indirectly by the conserved aspartate residue of the Walker-B motif. Figure 2.4b illustrates the Walker-A and Walker-B motifs and metal coordinating residues in UMP/CMP kinase.

Figure 2.4: The P-loop Kinase and PEPCK Families

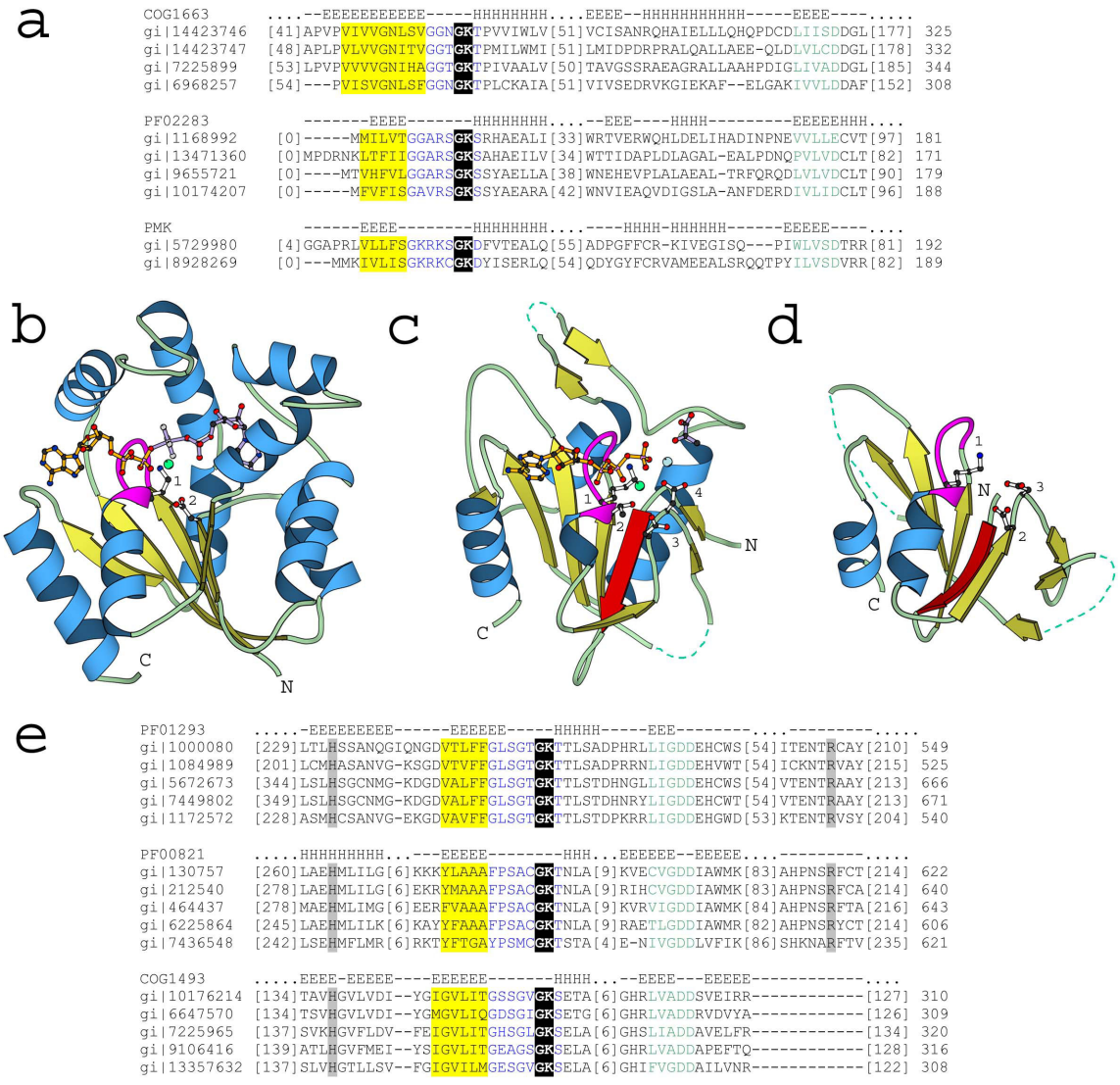


Figure 2.4: The P-loop Kinase and PEPCK Families. a) Alignment of representative sequences of phosphomevalonate kinase and the 2 Pfam members with non-trivial links to the P-loop kinase family. Blue and green letters signify the Walker-A and Walker-B motifs, respectively. Other details pertaining to alignment layout are described in Figure 2.1 legend. b) UMP/CMP kinase (pdb|1qf9) (Schlichting and Reinstein 1999) of the P-loop kinase family. Residues 1 and 2 are the conserved Lys19 of the Walker-A motif and the conserved Asp89 of the Walker-B motif, respectively. Lys19 coordinates the β- and γ-phosphates of ATP while Asp89 coordinates the Mg^{2+} cation. The P-loop is shown in magenta. In panels b and c, the nucleotide is orange, the phosphate-accepting substrate is purple, and the Mg^{2+} cation is a green ball. c) Phosphoenolpyruvate carboxykinase (pdb|1aq2) (Tari, Matte et al. 1997) of the PEPCK family. Residues 1 and 2 are Lys254 and Thr255 of the Walker-A motif, respectively. Residues 3 and 4 are the pair of conserved Asp residues (Asp268 and Asp269) in the PEPCK Walker-B motif. Lys254 coordinates the β- and γ-phosphates of ATP while Thr255, Asp268, and Asp269 coordinate the Mg^{2+} cation. The antiparallel β-strand is shown in red, and the P-loop is shown in magenta. The Mn^{2+} cation is a light blue ball. Most of the N-terminal domain and some elements of the C-terminal domain were removed for clarity. Dashed lines indicate insertions. d) HPr kinase/phosphatase (pdb|1jb1) (Fieulaine, Morera et al. 2001) of the PEPCK family. Residue 1 is Lys161 of the Walker-A motif and is involved in coordination of the β- and γ-phosphates of ATP. Residues 2 and 3 are the pair of conserved Asp residues (Asp178 and Asp179) in the Walker-B motif and are involved in coordinating the Mg^{2+} cation. The antiparallel β-strand is shown in red, and the P-loop is shown in magenta. Only the core of the C-terminal domain is shown. Dashed lines indicate insertions or disordered regions. e) The alignment of three Pfam members of the PEPCK family: PF01293 (phosphoenolpyruvate carboxykinase - ATP), PF00821 (phosphoenolpyruvate carboxykinase - GTP), and COG1493 (HPr kinase/phosphatase). The conserved histidine and arginine residues are believed to be involved in substrate binding and are catalytically important in PEPCK (Matte, Goldie et al. 1996). Blue and green letters signify the Walker-A and Walker-B motifs, respectively. Other details pertaining to alignment layout are described in Figure 2.1 legend.

The Walker A and Walker B motifs are common in nucleotide-binding proteins. In SCOP (version 1.55), the P-loop containing nucleotide triphosphate hydrolases fold contained 82 proteins in 15 families. Of these 82 proteins, 13 were kinases. Thus, ~84% of P-loop containing proteins with known structures were not kinases. The non-kinase P-loop containing proteins are mostly NTPases, which are likely to have come before the kinases since they catalyze simpler reactions and are involved in more fundamental biological processes.

A variety of catalytic mechanisms are utilized by the kinases in this family. For example, an iso-random Bi-Bi mechanism has been suggested for adenylate kinase (Sheng, Li et al. 1999). A mechanism involving the synchronous shift of a proton to the transferred phosphate group (similar to that proposed for protein kinases) has been suggested for the UMP/CMP-kinase reaction (Hutter and Helms 2000). If correct, such a mechanism could also apply to the other nucleotide-phosphorylating kinases in this family. However, the phosphorylation of some metabolites appears to require a base

catalyst (Miziorko 2000). This would apply to certain members of the P-loop kinase family including, for example, phosphoribulokinase and shikimate kinase.

Family 2b: phosphoenolpyruvate carboxykinase

The phosphoenolpyruvate carboxykinase (PEPCK) family consists of 3 Pfam/COG members: PF01293 (phosphoenolpyruvate carboxykinase - ATP), PF00821 (phosphoenolpyruvate carboxykinase - GTP), and COG1493 (HPr kinase/phosphatase). Members of this family are distinguished by their shared nucleotide-binding region, characterized by the topology of the β-sheet and the placement of the Walker-A motif and an atypical Walker-B motif within the nucleotide-binding fold. Solved structures of representatives from PF01293 and COG1493 illustrate this shared topology (Figure 2.4c,d). Although no structure representative of PF00821 was available at the time of the initial classification, members of this family were known to contain the characteristic Walker-A and Walker-B motifs. Similar predicted secondary structure distributions and conserved potential catalytic residues indicated that proteins from PF00821 and PF01293 share similar fold and active site architecture (Figure 2.4e). The solved structure of this kinase later confirmed this prediction.

PEPCK contains two α/β domains. The nucleotide-binding fold is located in the C-terminal domain and is composed of the 6-stranded mixed β-sheet and the surrounding α-helices (Matte, Goldie et al. 1996). The PEPCK family proteins contain the typical Walker-A motif and a deviant Walker-B motif (Figure 2.4e). Figures 2.4b and 2.4c illustrate the phosphate-binding loops of a P-loop kinase and PEPCK, respectively. Note the similar structures of the Walker-A motifs (in magenta) and the different spatial locations of the Walker-B aspartate residues between the two proteins. The topology of the nucleotide-binding fold of PEPCK differs from that in P-loop kinases. The central β-sheet of the PEPCK nucleotide-binding C-terminal domain is a mixed 6-stranded β-sheet with strand order 312456, and β-strands 1 and 5 are antiparallel to the rest of the sheet (Matte, Goldie et al. 1996), whereas the central β-sheet of P-loop kinases typically consists of 5 parallel β-strands of strand order 21345. Furthermore, while the β-strand

preceding the Walker-A motif and the β-strand preceding Walker-B motif are neighboring structural elements in space in P-loop kinases, PEPCK has an antiparallel β-strand that lies between the two β-strands associated with the Walker-A and Walker-B motifs. This β-strand is colored red in Figures 2.4c and 2.4d.

The C-terminal domain of HPr kinase/phosphatase (HPrK/P) is another member of the PEPCK family. The first solved structure of this domain is shown in Figure 2.4d, although due to poor electron density, only 5 β-strands in the core of this domain were visible (Fieulaine, Morera et al. 2001). Later structures showed that the C-terminal domain of HPrK/P also contained an additional β-strand in one of the disordered regions (Allen, Steinhauer et al. 2003), revealing that the topology of this β-sheet in HPrK/P is identical to that of the corresponding β-sheet in PEPCK. Additionally, the placement of the Walker-A and Walker-B motifs is similar in both HPrK/P and PEPCK (Figure 2.4e). In both kinases, the Walker-A motif is located on a βα loop following β-strand 2. The Walker-B motif is found at the C-terminal end of β-strand 3.

Two divalent metal cations are present in the active site of PEPCK (Figure 2.4c). The Mg^{2+} cation is coordinated by the threonine hydroxyl of the Walker-A motif and indirectly by the conserved pair of aspartate residues in the Walker-B motif. The Mn^{2+} cation is coordinated by the side chain nitrogens of a histidine and a lysine residue in addition to the second aspartate residue of the Walker-B motif. The Mg^{2+} cation interacts with oxygens of the β- and γ-phosphoryl groups of ATP while the Mn^{2+} cation associates with an oxygen of the γ-phosphoryl group and hypothetically with the enolate oxygen of pyruvate during catalysis (Tari, Matte et al. 1997). PEPCK is different from other kinases in that this enzyme catalyzes both the decarboxylation and phosphorylation of its substrate, oxaloacetate, to form phosphoenolpyruvate. However, the details of this mechanism are yet unknown.

Family 2c: phosphoglycerate kinase

Phosphoglycerate kinase is composed of two α/β/α domains. Both domains adopt a Rossmann-like fold and each contains a 6-stranded parallel β-sheet. The N-terminal

domain β-sheet has strand order 342156 while the C-terminal domain β-sheet has strand order 321456. The active site of this enzyme is located in the cleft between these two domains (Watson, Walker et al. 1982). In this enzyme, the C-terminal domain binds ATP while the N-terminal domain binds the 3-phosphoglycerate substrate. The nucleotide binds at the edge of the β-sheet in the C-terminal domain and is roughly perpendicular to the β-strands. There are no sequence or structural motifs resembling the Walker-A or Walker-B motifs in phosphoglycerate kinase.

The primary factor contributing to catalysis in phosphoglycerate kinase is transition state stabilization. In this enzyme, all three peripheral oxygens of the transferred phosphate are stabilized by interactions with protein residues or the required divalent cation in the transition state. However, only two of these oxygens are stabilized when the phosphate is fully bonded to either the phosphate donor (ATP) or acceptor (phosphoglycerate substrate) (Bernstein and Hol 1998).

Family 2d: aspartokinase

This family contains only one Pfam member (PF00696: Amino acid kinase family) and includes aspartokinase, carbamate kinase, acetylglutamate kinase, glutamate 5-kinase, and uridylate kinase. The only kinase in this family with a solved structure at the time of initial survey was carbamate kinase from *Enterococcus faecalis* (Marina, Alzari et al. 1999) and *Pyrococcus furiosus* (Ramon-Maiques, Marina et al. 2000), although structures of acetylglutamate kinase from several species are now available as well. The nucleotide-binding domain in aspartokinases is composed of three layers (α/β/α), including a β-sheet with Rossmann-like topology (Marina, Alzari et al. 1999). The bound nucleotide is located at the edge of the β-sheet strands and is approximately perpendicular to the direction of the β-strands as is shown in the complex structure of carbamate kinase-like carbamoyl phosphate synthetase from *P. furiosus* (Ramon-Maiques, Marina et al. 2000) and acetylglutamate kinase from *Escherichia coli* (Gil-Ortiz, Ramon-Maiques et al. 2003).

Family 2e: phosphofructokinase-like

The phosphofructokinase-like family contains 4 Pfam members. Links between 3 of these members are trivial via PSI-BLAST. The link between another Pfam member, PF01219 (Prokaryotic diacylglycerol kinase), to the phosphofructokinase-like family was established through multiple sequence alignment analysis and secondary structure predictions (Figure 2.5a). The fold of phosphofructokinase consists of two α/β/α domains. The nucleotide-binding N-terminal domain has a 7-stranded mixed β-sheet of strand order 3214567, where β-strands 3 and 7 are antiparallel to the rest of the β-sheet (Figure 2.5b). The active site is located in the cleft between the two domains (Shirakihara and Evans 1988). The ATP, which is positioned above the β-sheet of the N-terminal domain and is approximately parallel to the β-strands, sits between an α-helix and a long loop segment (Figure 2.5b).

Family 2f: ribokinase-like

The ribokinase-like family contains several carbohydrate kinases. All 3 Pfam members in this family have trivial PSI-BLAST links. An additional grouping of archaeal phosphofructokinase/glucokinase sequences can also be placed in the ribokinase-like family. An alignment for this assignment is shown in Figure 2.5d. The solved structure of glucokinase from *Thermococcus litoralis* shows that the structure of this archaeal ADP-dependent kinase is indeed similar to the other members of the ribokinase-like family (Ito, Fushinobu et al. 2001).

The core of the ribokinase-like fold is a 3-layered domain (α/β/α) with a central 8-stranded β-sheet. The strand order of this β-sheet is 21345678 with β-strand 7 antiparallel to the rest of the β-sheet. Ribokinase also has an additional subdomain composed of a 4-stranded β-sheet. In addition to acting as a "substrate lid", this β-sheet forms the dimerization surface (Sigrell, Cameron et al. 1998). In ribokinase, the nucleotide-binding site is found along a shallow groove in the core of the sole α/β/α domain (Sigrell, Cameron et al. 1998). The ATP moiety lies at the edge of the β-sheet and is roughly perpendicular to the adjacent β-strands (Figure 2.5c).

Figure 2.5: The Phosphofructokinase-Like and Ribokinase-Like Families

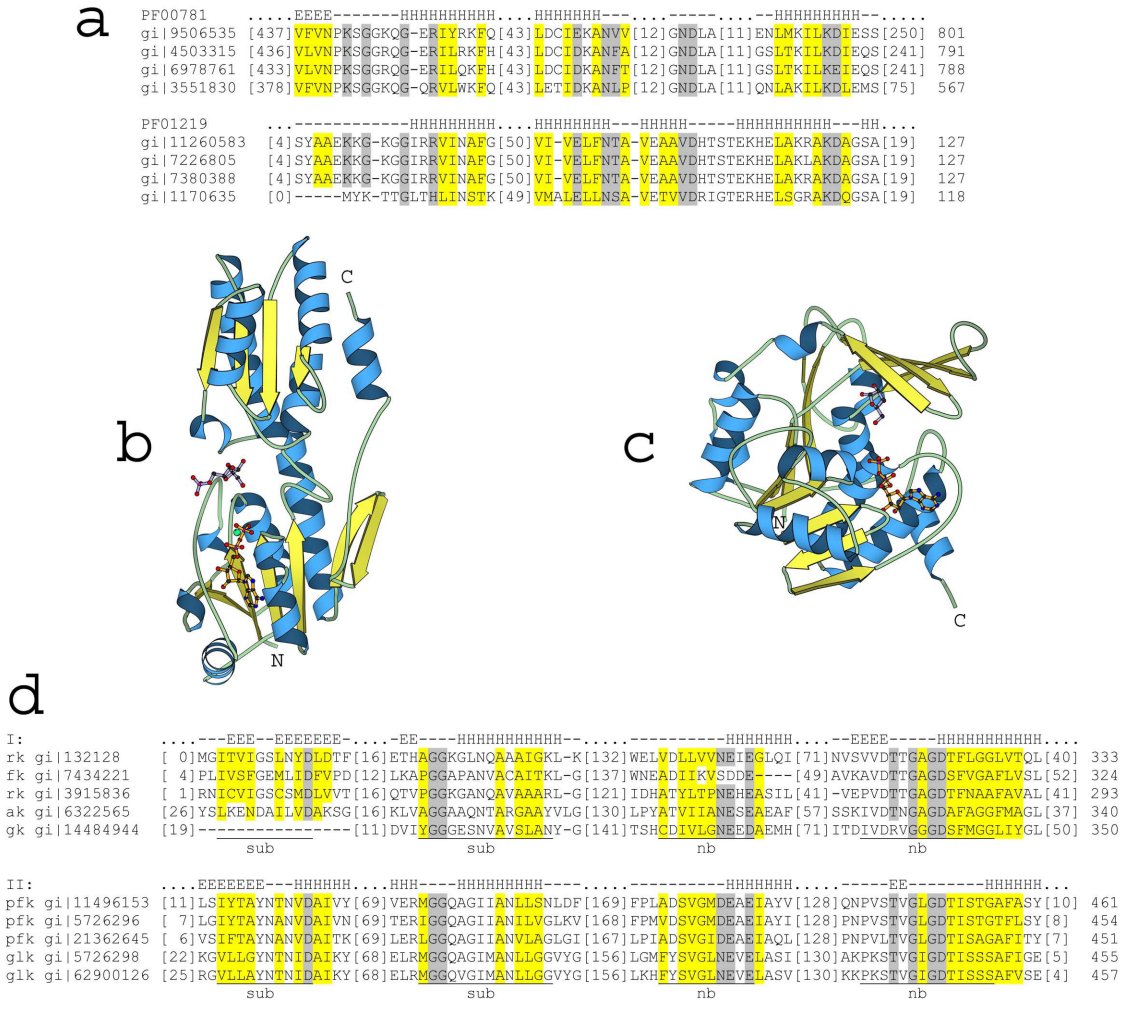


Figure 2.5: The Phosphofructokinase-like and Ribokinase-Like Families. a) Alignment of representative sequences from the two Pfam members containing diacylglycerol kinase in the phosphofructokinase-like family: PF00781 (Diacylglycerol kinase catalytic domain - presumed) and PF01219 (Prokaryotic diacylglycerol kinase). Details pertaining to alignment layout are described in Figure 2.1 legend. b) Phosphofructokinase (pdb|4pfk) (Evans, Farrants et al. 1981). In panels b and c, the nucleotide is orange, the phosphate-accepting substrate is purple, and the Mg^{2+} cation is a green ball. c) Ribokinase (pdb|1rkd) (Sigrell, Cameron et al. 1998). d) Addition of distant members of the ribokinase-like family. I: representative sequences of ribokinase-like family: ribokinase (rk), fructokinase (fk), adenosine kinase (ak), gluconate kinase (gk); II: representative sequences of archaeal phosphofructokinase/glucokinase: phosphofructokinase (pfk), glucokinase (glk). The underlined residues are conserved motifs in the nucleotide-binding pocket (nb) and the substrate binding pocket (sub) (Carret, Delbecq et al. 1999). Other details pertaining to alignment layout are described in Figure 2.1 legend.

Family 2g (removed in updated survey): L-2-haloacid dehalogenase (HAD-like)

The HAD-like family contained only 2 putative kinase sequence members, which are annotated as bifunctional homoserine kinase/phosphoserine phosphatase (ThrH) from *Pseudomonas aeruginosa* (Patte, Clepet et al. 1999). PSI-BLAST establishes a link between these sequences and phosphoserine phosphatase, which has the HAD-like fold (Cho, Wang et al. 2001; Wang, Kim et al. 2001). The core of the HAD-like fold is a 6-stranded parallel β-sheet of strand order 321456 with an additional 4-helix bundle subdomain (Hisano, Hata et al. 1996). The nucleotide-binding site in HAD-like enzymes is suggested to be located in the cleft between the two domains of the fold (Hisano, Hata et al. 1996), however the details of the orientation of ATP are unknown.

HAD-like family enzymes typically utilize a conserved aspartate residue for nucleophilic catalysis in a reaction that includes a covalent intermediate involving the phosphate group and the aspartate residue (Collet, Stroobant et al. 1999; Morais, Zhang et al. 2000). This mechanism could potentially be used for a kinase phosphotransfer reaction as well. Alignment of the homoserine kinase isozyme with phosphoserine phosphatase suggested that Asp7 (gi|4138297) is the residue that is phosphorylated for the formation of the covalent intermediate in the homoserine kinase isozyme.

Although these sequences were annotated as kinases and do carry out a phosphotransfer reaction, they do not use ATP as the phosphate donor and therefore do not meet the definition of kinase used in this study. Earlier genetic studies of the *thrH* gene of *Pseudomonas aeruginosa* have shown that over-expression of ThrH complements both homoserine kinase and phosphoserine phosphatase activities *in vivo*. The gene product of *thrH* was thus annotated as "bifunctional homoserine kinase/phosphoserine phosphatase isoenzyme" (Patte, Clepet et al. 1999). A more recent structural and biochemical study of ThrH has shown that this protein does not have ATP-dependent kinase activity. Instead it possesses phosphoserine phosphatase activity and is also able to transfer a phosphate group from phosphoserine to homoserine, presumably via a phospho-enzyme intermediate (Singh, Yang et al. 2004). Thus, although ThrH is able to generate phosphohomoserine and complement homoserine kinase activity *in vivo*, it

achieves this through a completely different chemical mechanism from that of true homoserine kinase. Thus, ThrH is in fact a phosphatase and phosphotransferase but not a kinase and is subsequently removed from the kinase classification.

Family 2h: thiamin pyrophosphokinase

Thiamin pyrophosphokinase (TPK) is a two-domain protein. The ATP binding C-terminal domain has Rossmann-like topology and is composed of three layers (α/β/α) with a central 6-stranded parallel β-sheet of strand order 432156, while the N-terminal domain consists of a 4-stranded β-sheet and a 6-stranded β-sheet which form a flattened sandwich ("jelly-roll topology") (Baker, Dorocke et al. 2001). TPK is a homodimer with the active site located in a cleft at the interface of the component monomers. Thus, active site residues are contributed by the N-terminal domain of one monomer and the C-terminal domain of the opposing monomer (Baker, Dorocke et al. 2001). The precise orientation of the nucleotide in the active site has not been determined.

Family 2i (previously Group 15): glycerate kinase

Glycerate kinase from *Neisseria meningitidis* (Rajashankar, Kniewel et al. *To be published*) is a member of a glycerate kinase family (previously Group 15) that did not have a structural representative in the initial kinase survey. The fold of this protein consists of two non-similar α/β domains (Figure 2.6a). The N-terminal domain contains a central 5-stranded parallel β-sheet (strand order 21345) and several surrounding helices with Rossmann-like topology. The C-terminal domain contains a central 6-stranded mixed β-sheet (strand order 123456), with β-strand 2 antiparallel to the rest of the β-sheet. In this structure, the C-terminal domain is inserted between β-strands 2 and 3 of the N-terminal domain. The active site is likely to be in the cleft between the two α/β domains. Although this structure does not include a bound nucleotide or substrate, the two sulfate groups observed in the presumed active site may indicate the locations of the nucleotide and glycerate binding sites. In this structure, eight highly conserved polar or charged residues have side chains pointing into the presumed active site (Figure 2.6a).

Six of these eight residues, including two lysines, are contributed by the Rossmann-like domain. Furthermore, the crevice in which the sulfate group is located is significantly larger in the Rossmann-like domain that the corresponding crevice in the C-terminal domain, suggesting that the Rossmann-like domain may accommodate a larger molecule such as ATP. Based on the presence of these conserved lysine residues, which are common features of nucleotide binding sites in kinases, and the large crevice that may serve as the ATP-binding site, the Rossmann-like domain is predicted to perform the nucleotide binding role in this kinase. Thus, this family of glycerate kinases is incorporated into the Rossmann-like fold group as an additional family. Members of this

Figure 2.6: Two Glycerate Kinase Families

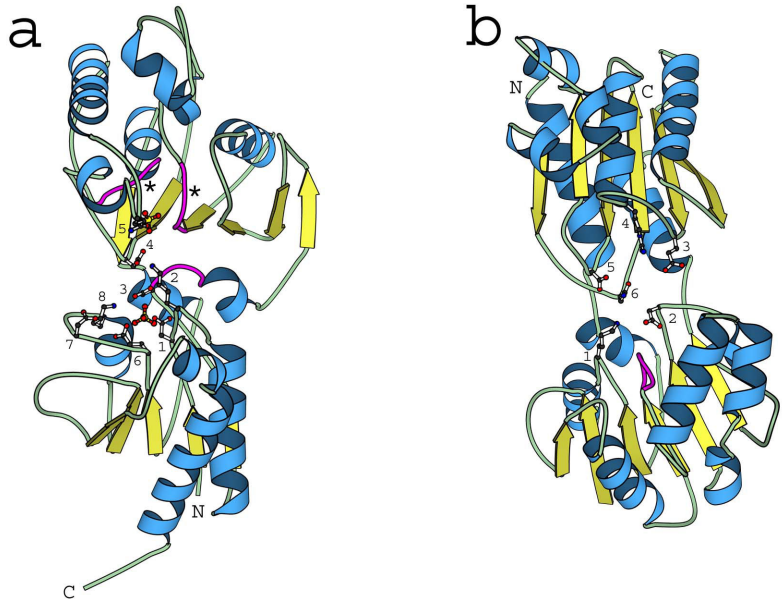


Figure 2.6: Two Glycerate Kinase Families. a) *Neisseria meningitides* glycerate kinase (pdb|1to6), a representative of first glycerate kinase family (Group 15 in initial survey; Family 2i in updated survey). Highly conserved amino acids with side chains pointing into the presumed active site include six residues from the Rossmann-like domain (**1** - Asp8, **2** - Lys11, **3** - Asp43, **6** - Glu286, **7** - Asp290, and **8** - Lys297) and two residues from the inserted domain (**4** - Asp191 and **5** - Gln209). Glycine rich loops are shown in magenta; those predicted as functionally important in the initial kinase classification are marked with asterisks. Sulfate groups are shown in ball-and-stick representation. b) *Thermotoga maritima* putative glycerate kinase (pdb|1o0u), a member of the second glycerate kinase family (Group 16 in initial survey, Group 10 in updated survey). Highly conserved amino acids with side chains pointing into the presumed active site include two residues from the Rossmann-like domain (**1** - Lys47 and **2** - Asp189) and four residues from the C-terminal domain (**3** - Glu312, **4** - Arg325, **5** - Asp351, and **6** - Asn407). The glycine rich loop is shown in magenta.

family of glycerate kinases are from bacterial species, primarily of the *firmicutes* group and of the gamma subdivision of the proteobacterial group. It should be noted that this family is distinct from a second family of putative glycerate kinases (Group 16 in initial survey, Group 10 in updated survey), which consists of proteins from eukaryotes and archaea in addition to several different bacterial species.

2.4.3 Group 3: ferredoxin-like fold kinases

Group 3 is composed of kinases whose nucleotide-binding domain core resembles the ferredoxin-like fold. The four families in this group are nucleoside-diphosphate (NDP) kinase, 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK), guanido kinases, and histidine kinase. The ferredoxin fold (also known as the α-β plait fold or the α+β sandwich) is characterized by its βαββαβ unit with the 4-stranded antiparallel β-sheet having strand order 2314 and two α-helices on one side of the β-sheet. One exception is the histidine kinase family, in which the core of the nucleotide-binding domain is related to the ferredoxin-like topology by circular permutation. An interesting feature of this group is that the mode of nucleotide binding differs significantly between each of the families, not only in terms of structural elements utilized in nucleotide-protein interactions, but also in the orientation and position of the nucleotide relative to the ferredoxin-like core. Figure 2.7 illustrates the orientation of the bound nucleotide in NDP kinase, HPPK, arginine kinase, and histidine kinase.

Family 3a: NDP kinase

In addition to the ferredoxin-like core of NDP kinase, this enzyme also has several other secondary structure elements, including the Kpn loop, an α-helical hairpin, and a C-terminal extension (Figure 2.7a). The Kpn loop is a small, compact structural element containing an interesting combination of helical structures: a turn of 3_{10} helix, a turn of polyproline II left-handed helix, and a turn of standard α-helix (Janin, Dumas et al. 2000). The Kpn loop and the α-helical hairpin constitute a nucleotide-binding site that

Figure 2.7: Nucleotide Binding in the Ferredoxin-Like Kinase Group

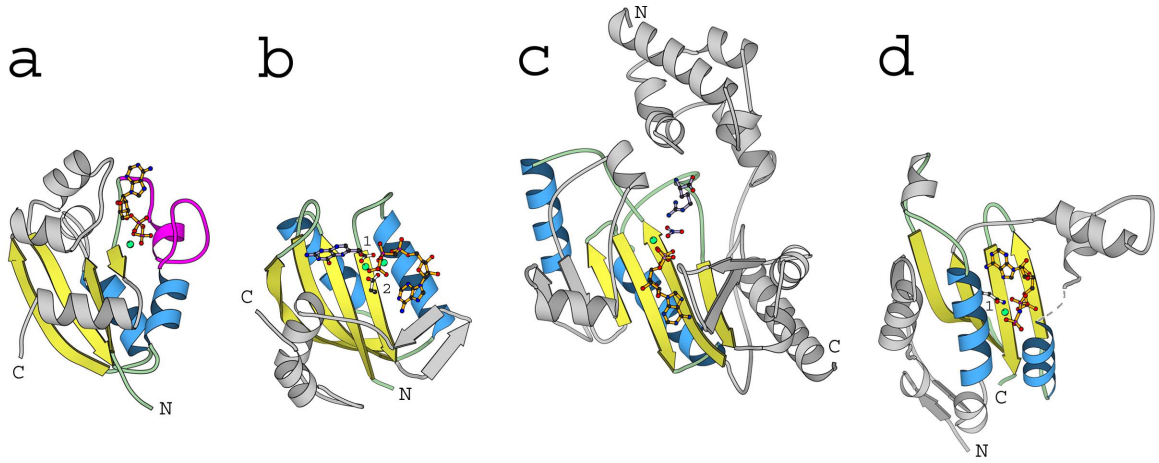


Figure 2.7: Nucleotide Binding in the Ferredoxin-Like Kinase Group. a) Nucleoside-diphosphate kinase (pdb|2bef) (Xu, Morera et al. 1997) of the NDP kinase family. The Kpn loop is shown in magenta. b) 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (pdb|1eqo) (Blaszczyk, Shi et al. 2000) of the HPPK family. Residues 1 (Asp95) and 2 (Asp97) are involved in Mg^{2+} coordination. c) Arginine kinase (pdb|1bg0) (Zhou, Somasundaram et al. 1998) of the guanido kinase family. d) Histidine kinase (pdb|1i59) (Bilwes, Quezada et al. 2001) of the histidine kinase family. Residue 1 (Asn409) is involved in the coordination the Mg^{2+} cation. Dashed lines indicate disordered regions. In these structures, the ferredoxin-like core of the proteins is shown in yellow β-strands and blue α-helices. Any additional secondary structure elements are shown in grey. The nucleotide is orange, the phosphate-accepting substrate is purple, and Mg^{2+} cations are green balls.

is unique to NDP kinase (Figure 2.7a). In terms of the orientation and position of this substrate, the nucleotide lies at the edge of the β-sheet that defines the ferredoxin-like core, with the adenine base more distant from the β-sheet and the phosphate tail extending towards the β-sheet (Morera, Lascu et al. 1994).

Catalysis in NDP kinase occurs by a ping-pong mechanism in which the phosphoryl group is transferred first from the nucleoside triphosphate to the enzyme, involving the formation of a covalent intermediate of the phosphate with a histidine residue in the active site, followed by the transfer of the phosphate to the nucleoside diphosphate substrate.

Family 3b: HPPK

In addition to the ferredoxin-like core of HPPK, this enzyme has a pair of β-strands and a pair of α-helices at the C-terminal end of the protein (Stammers, Achari et

al. 1999; Blaszczyk, Shi et al. 2000). In HPPK, ATP is bound in the region between the α2-β4 connecting loop and a short C-terminal β-strand that is not part of the ferredoxin-like core (Figure 2.7b). The ATP lies at the edge of the β-sheet of the ferredoxin-like core and is curled such that the adenine base and ribose sugar angle away from the β-sheet while the triphosphate tail points towards the β-sheet. The adenine base and the ribose sugar lie in the same plane as the β-sheet. However, the triphosphate tail reaches over the β-sheet on the opposite side as the α-helices associated with the ferredoxin-like core (Stammers, Achari et al. 1999; Blaszczyk, Shi et al. 2000).

HPPK utilizes two Mg^{2+} cations in its active site, each of which is coordinated by two aspartate residues. The mechanism of HPPK has not yet been elucidated, but it is presumed to be a direct in-line transfer of the pyrophosphate group. Suggestions for the mechanism include roles for the two Mg^{2+} cations and acid-base catalysis with a water molecule acting as the general base (Blaszczyk, Shi et al. 2000).

Family 3c: guanido kinases

The fold of the guanido kinase family consists of two domains. The smaller N-terminal domain is composed entirely of α-helices, and the nucleotide-binding C-terminal domain is composed of an 8-stranded β-sheet of strand order 23451687 which is flanked by 7 α-helices (Fritz-Wolf, Schnyder et al. 1996; Zhou, Somasundaram et al. 1998). The middle 4 β-strands of the β-sheet and associated α-helices have ferredoxin-like fold topology. Figure 2.7c illustrates the ferredoxin-like topology within this fold.

In arginine kinase, ATP binding is accomplished by interactions with 5 arginine residues and a Mg^{2+} cation. The ATP lies in the plane above the β-sheet, rather than at the edge of the β-sheet, and is oriented approximately parallel to the β-strands (Figure 2.7c). The nucleotide is positioned above the center two β-strands of the 4-stranded section that resembles the ferredoxin-like fold. In this enzyme, the bound nucleotide and the α-helices that compose the ferredoxin-like topology lie on opposite sides of the β-sheet (Zhou, Somasundaram et al. 1998).

Arginine kinase catalyzes an associative in-line phosphotransfer reaction. The primary factor in catalysis appears to be substrate alignment by positioning reaction components in close proximity and promoting proper alignment of orbitals, although acid-base catalysis, polarization, and transition state stabilization may also contribute to the reaction (Zhou, Somasundaram et al. 1998).

Family 3d: Histidine kinases

There are 2 Pfam/COG members of this family. The link between the members is trivial via PSI-BLAST. Histidine kinases (HKs) catalyze a trans-autophosphorylation reaction in the two-component system of signal transduction. The fold of the ATP-domain (the catalytic domain) is an α/β sandwich composed of a 5-stranded β-sheet and 4 α-helices (Figure 2.7d) (Tanaka, Saha et al. 1998; Bilwes, Quezada et al. 2001). The core of this fold has a tertiary structure similar to that of the ferredoxin fold. The HK fold has topology αββαββ with strand order 3421, which can be related to the ferredoxin-like topology by a circular permutation consisting of cutting the loop between the first and second β-strands and connecting the natural termini.

There are two classes of HKs, which can be differentiated by their domain organization (Dutta, Qin et al. 1999). EnvZ is a member of class I in which the H-domain (the domain that contains the histidine phosphorylation site) directly precedes the ATP-domain. CheA is a member of class II in which the H-domain and ATP-domain are separated by at least one other domain. The ATP-domains of HKs are structurally similar to the ATP-binding domains of the GHL family (DNA gyrase/Hsp90/MutL) (Tanaka, Saha et al. 1998; Bilwes, Quezada et al. 2001). There is currently some debate as to whether both HK classes or only class II bind ATP in the same conformation as the GHL family (Tanaka, Saha et al. 1998; Bilwes, Quezada et al. 2001). In both classes, there are four conserved motifs that contribute to the nucleotide-binding pocket, namely the N, G1, F, and G2 boxes (Robinson, Buckler et al. 2000). The required Mg^{2+} cation is coordinated by direct interactions with an asparagine and indirectly by a histidine and an arginine (Bilwes, Quezada et al. 2001).

In HK, the ATP lies above the β-sheet that is part of the ferredoxin-like core (Figure 2.7d). The nucleotide is above two of the β-strands and is found at the edge of the β-sheet rather than the center of it. In HK, the adenine base is nearer to the β-sheet than the triphosphate tail which angles away from the β-strands. The nucleotide and the α-helices associated with the ferredoxin-like core are located on the same side of the β-sheet (Bilwes, Quezada et al. 2001).

HK, as noted above, operates via the two-component system. Here, one HK monomer phosphorylates a histidine residue in the other monomer of the homodimer, which results in a high energy phosphoryl group. The regulatory domain of the cognate response regulator (RR) then catalyzes the reaction that transfers the phosphoryl group from the histidine residue to an aspartate residue in the RR. The mechanism is somewhat similar to that of NDP kinase in that both involve the formation of a high energy phospho-histidine residue. However, NDP kinase phosphorylates a histidine residue in its own active site while HK phosphorylates a histidine in another HK monomer.

2.4.4 Group 4: ribonuclease H-like kinases

There are 4 Pfam/COG members in this group. Three of these members have trivial links via PSI-BLAST. The fourth member (PF00871: Acetokinase family, which contains acetate kinase and butyrate kinase) was predicted to be a member of this family by Buss *et al* (Buss, Ingram-Smith et al. 1997). The solved structure of acetate kinase shows that this enzyme does in fact adopt the ribonuclease H-like fold (Buss, Cooper et al. 2001). Multiple alignment of 2-dehydro-3-deoxygalactonokinase sequences indicates that this kinase activity also belongs to the ribonuclease H-like group (Figure 2.8a). The ribonuclease H-like group contains the ASKHA (acetate and sugar kinase/hsc70/actin) superfamily, the structures of which are characterized by duplicate domains of the ribonuclease H-like fold. The ribonuclease H-like fold is composed of three layers (α/β/α) (Figure 2.8b). The 5-stranded mixed β-sheet has strand order 32145, with β-strand 2 antiparallel to the rest of the β-sheet. The topology of the core of this fold is

Figure 2.8: The Ribonuclease H-Like Family, TIM β/α-Barrel Kinase Family, and GHMP Kinase Family

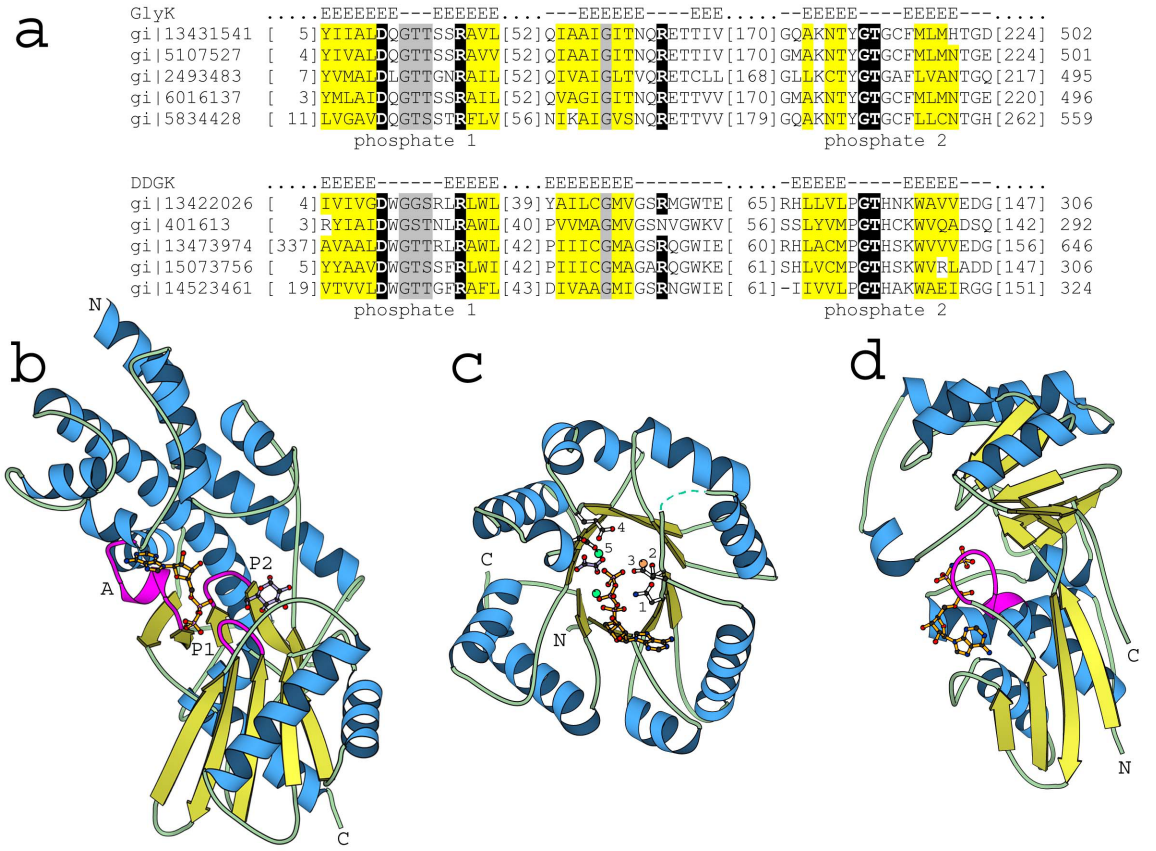


Figure 2.8: The Ribonuclease H-Like Family, TIM β/α-Barrel Kinase Family, and GHMP Kinase Family. a) Alignment of 2-dehydro-3-deoxygalactonokinase (DDGK) sequences and glycerol kinase (GlyK) sequences of the ribonuclease H-like family. The PHOSPHATE 1 and PHOSPHATE 2 motifs are indicated. Other details pertaining to alignment layout are described in Figure 2.1 legend. b) Hexokinase (pdb|1dgk) (Aleshin, Kirby et al. 2000) of the ribonuclease H-like family. The loop regions of the conserved nucleotide-binding motifs [PHOSPHATE 1 (P1), PHOSPHATE 2 (P2), ADENOSINE (A)] are shown in magenta. Only the C-terminal domain is shown. c) Metal cofactor coordination and nucleotide orientation in the TIM β/α-barrel kinase family (pyruvate kinase, pdb|1a49 (Larsen, Benning et al. 1998)). Residues 1 (Asn74), 2 (Ser76), and 3 (Asp112) coordinate the K^+ cation. Residues 4 (Glu271) and 5 (Asp295) coordinate one of the Mg^{2+} cations. The C-terminal subdomain was removed for clarity; dashed lines indicate the location of this insertion. d) Homoserine kinase (pdb|1h72) (Krishna, Zhou et al. 2001) of the GHMP kinase group. The novel P-loop is shown in magenta. In these structures, the nucleotide is orange, the phosphate-accepting substrate is purple, Mg^{2+} cations are green balls, and K^+ cations are orange balls.

βββαβαβα. Nucleotide binding and divalent metal coordination are achieved by interactions of ATP with several motifs conserved within the ASKHA superfamily (Bork, Sander et al. 1992). These conserved motifs include the ADENOSINE motif that interacts with ribosyl and the α-phosphoryl group of ATP, the PHOSPHATE 1 motif that interacts with Mg^{2+} through coordinated water molecules, and the PHOSPHATE 2 motif that interacts with the β- and γ-phosphoryl groups of ATP (Figure 2.8b). A modeled active site of hexokinase predicts that the required divalent metal cation is not directly liganded to the protein, but is positioned by coordinated water molecules (Aleshin, Kirby et al. 2000). The presumed mechanism of kinases in this group is acid-base catalysis. In hexokinase, an aspartate residue is the putative catalytic base that deprotonates the 6-hydroxyl of glucose (Arora, Filburn et al. 1991; Aleshin, Kirby et al. 2000).

The structure of eukaryotic pantothenate kinase (Group 14 in initial survey) has not been solved experimentally. The crystal structure of prokaryotic pantothenate kinase identifies this enzyme as a member of the P-loop kinase family (Yun, Park et al. 2000). However, due to the lack of sequence identity between the prokaryotic and eukaryotic versions of this protein (Calder, Williams et al. 1999) in conjunction with dissimilar predicted secondary structure patterns, eukaryotic pantothenate kinase is expected to adopt a fold distinct from its prokaryotic counterpart. Although standard sequence similarity search methods failed to obtain any reasonable structural assignment, several fold recognition servers strongly suggested that the eukaryotic pantothenate kinases adopt a duplication of the ribonuclease H-like fold, although the closest structural template identified by 3D-Jury is a non-kinase homolog of this family (2-hydroxyglutaryl-CoA dehydratase component A; pdb|1hux (Locher, Hans et al. 2001)). The conservation of the functionally important ADENOSINE, PHOSPHATE 1, and PHOSPHATE 2 motifs in the eukaryotic pantothenate kinases is noted in Figure 2.9. Thus, based on the presence of these distinguishing motifs in addition to the similarity of secondary structure patterns (Figure 2.9), the eukaryotic pantothenate kinases are included with the ribonuclease H-like family in the updated survey.

Figure 2.9: Eukaryotic Pantothenate Kinase is a Ribonuclease H-Like Kinase

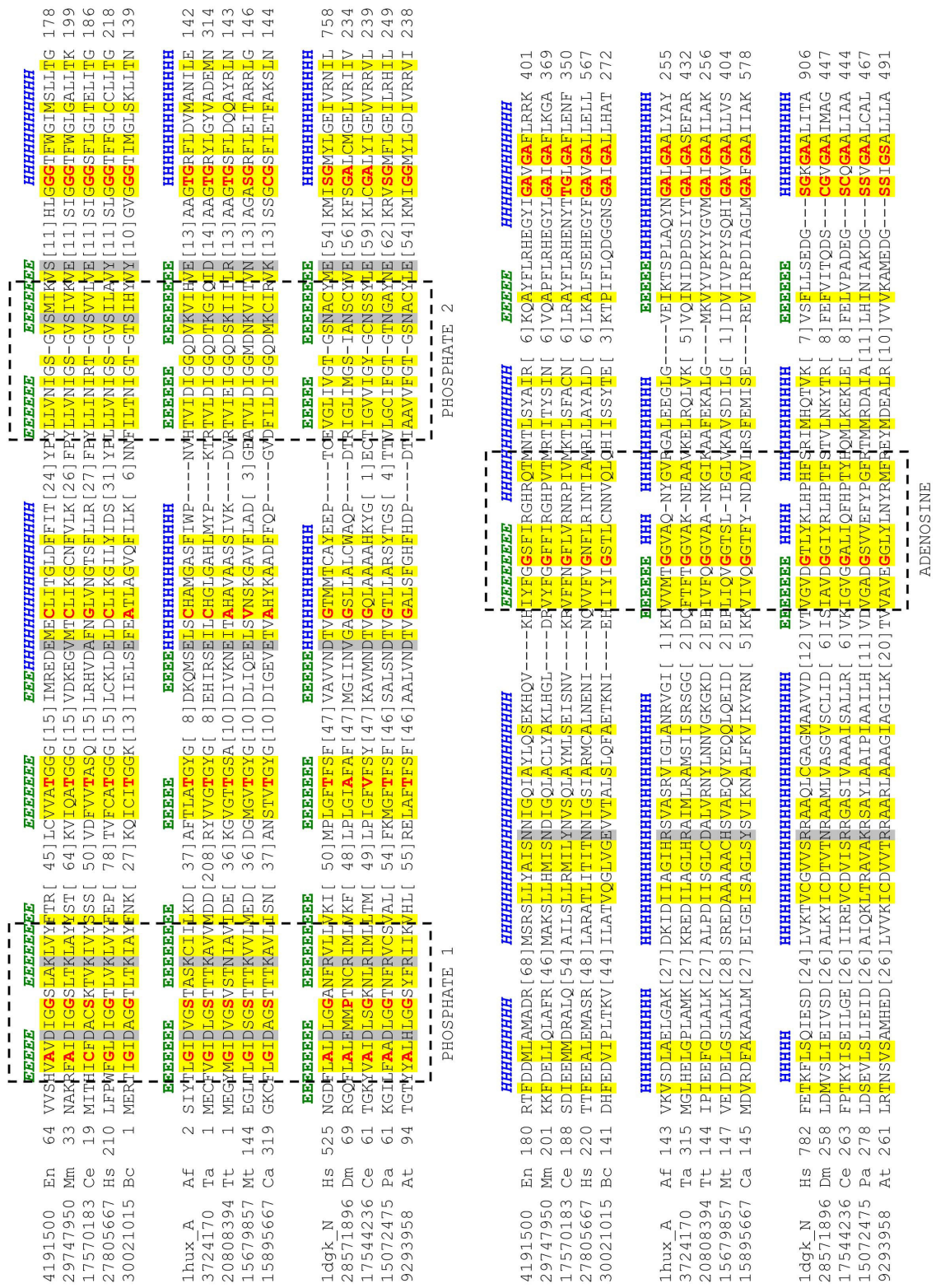


Figure 2.9: Eukaryotic Pantothenate Kinase is a Ribonuclease H-Like Kinase. Alignment for representative sequences of the eukaryotic pantothenate kinase family and two related ribonuclease H-like families with known structure: 2-hydroxyglutaryl-CoA dehydratase component A (pdb|1hux (Locher, Hans et al. 2001)) and hexokinase I (pdb|1dgk (Aleshin, Kirby et al. 2000)). The PHOSPHATE 1, PHOSPHATE 2, and ADENOSINE motifs are indicated by dashed boxes. Sequences are labeled according to the NCBI gi number or PDB code and an abbreviation of the species name. First and last residue numbers are indicated before and after each sequence. Numbers of excluded residues are specified in square brackets. Residue conservation is denoted with the following scheme: mostly hydrophobic positions, highlighted yellow; mostly charged/polar positions, highlighted grey; small residues, red bold text. Locations of predicted (gi|4191500) and observed (pdb|1hux, pdb|1dgk) secondary structure elements (E, β-strand; H, α-helix) are marked above the sequences in italics and normal font, respectively. Abbreviations of species names are as follows: Af *Acidaminococcus fermentans*, At *Arabidopsis thaliana*, Bc *Bacillus cereus*, Ca *Clostridium acetobutylicum*, Ce *Caenorhabditis elegans*, Dm *Drosophila melanogaster*, En *Emericella nidulans*, Hs *Homo sapiens*, Mm *Mus musculus*, Mt *Methanothermobacter thermautotrophicus*, Pa *Pichia angusta*, Ta *Thauera aromatica*, Tt *Thermoanaerobacter tengcongensis*.

2.4.5 Group 5: TIM β/α-barrel fold kinases

Group 5 is described by the TIM β/α-barrel fold, which consists of an 8-fold repeat of βα units that form a closed barrel. The barrel is composed of an inner layer of 8 parallel β-strands of strand order 12345678 and an outer layer of 8 α-helices (Figure 2.8c). This fold characterizes many different enzyme families that have extremely low sequence similarity and catalyze unrelated reactions (Hegyi and Gerstein 1999; Wierenga 2001). The active site of all TIM β/α-barrel enzymes is located at the C-terminal end of the parallel β-strands. The only kinase known to adopt this fold is pyruvate kinase, which also has an additional domain inserted at the C-terminal end of β-strand 3 (Larsen, Laughlin et al. 1994). The nucleotide-binding pattern of pyruvate kinase is thought to be novel (Larsen, Benning et al. 1998). The triphosphate tail of the ATP is held by hydrogen bond interactions with two arginines, an asparagine, a lysine, and three metal cations (two Mg^{2+} and one K^{+}). The adenine ring sits in a pocket that is bounded by a histidine, a proline, and a tyrosine residue. One distinctive feature of the pyruvate kinase active site is the coordination of each of the γ-phosphoryl peripheral oxygens to a different inorganic cofactor (Larsen, Benning et al. 1998). One of the Mg^{2+} cations is coordinated by the carboxylate groups of an aspartate and a glutamate residue. The other

Mg^{2+} cation is not directly liganded to the protein. The K^{+} cation is coordinated by the carbonyl of a threonine, the hydroxyl of a serine, the carboxylate of an aspartate, and the carboxyamide of an asparagine. The coordination of the metal cofactors is detailed in Figure 2.8c. The reaction catalyzed by pyruvate kinase is presumed to be direct in-line phosphotransfer via acid-base catalysis, although the group or groups responsible for the acid-base catalysis are yet to be identified. The possibility that a proton relay through a series of conserved residues may be responsible for acid-base catalysis has also been suggested (Larsen, Benning et al. 1998).

2.4.6 Group 6: GHMP kinases

The members of the GHMP kinase superfamily constitute Group 6. The GHMP kinase superfamily was named after the first initial of its original four members: **g**alactokinase, **h**omoserine kinase, **m**evalonate kinase, and **p**hosphomevalonate kinase (Bork, Sander et al. 1993), although several additional kinase activities were later included in this group. The crystal structure of only two members of this family, homoserine kinase and mevalonate kinase, were solved at the time of the initial kinase classification (Zhou, Daugherty et al. 2000; Yang, Shipman et al. 2002), although structures of several other GHMP superfamily members are now available. The fold of this group consists of two α+β domains, with the active site in the cleft between the two domains (Figure 2.8d). The N-terminal domain contains two β-sheets and 4 α-helices. The C-terminal domain has a ferredoxin-like core with four additional α-helices. The nucleotide-binding site resides mostly with the N-terminal domain. Nucleotide binding is accomplished by a novel P-loop with a conserved PXXXGSSAA motif. This feature performs a similar function to the P-loop of Family 2a, but the sequence motif is quite different. The structure of homoserine kinase revealed the presence of an unusual left-handed βαβα unit in the N-terminal domain. The second βα loop in this unit contains the novel phosphate binding loop (Figure 2.8d) (Zhou, Daugherty et al. 2000). Notably, the orientation of the ATP is different in the GHMP phosphate-binding loop than in the

classical Walker-A P-loop. A glutamate acts to coordinate the essential Mg^{2+} cation. In homoserine kinase, it has been suggested that the homoserine hydroxyl group is deprotonated not by a catalytic base, but by interaction with the γ-phosphate in a mechanism similar to that proposed for protein kinases (Krishna, Zhou et al. 2001). There are 4 Pfam/COG members of this group, and the links between these 4 members are trivial via PSI-BLAST.

2.4.7 Group 7: AIR synthetase (PurM-like) fold kinases

Two kinases, thiamine-phosphate kinase and selenide, water dikinase, belong to this group. These kinases adopt a structural fold similar to aminoimidazole ribonucleotide synthetase (AIR synthetase, PurM) (Li, Kappock et al. 1999). The Group 7 kinases have two α/β domains (Figure 2.10). The N-terminal domain contains a mixed 4-stranded β-sheet with 4 α-helices on one side of the β-sheet. The C-terminal domain has a mixed 6-stranded β-sheet flanked by 7 α-helices. Four of the β-strands and 2 of the α-helices in the C-terminal domain adopt a tertiary structure and topology that resembles the ferredoxin-like fold. The crystal structure indicates that this enzyme exists as a dimer, with the active site likely to be located in a cleft between the two subunits (Li, Kappock et al. 1999). A sulfate is bound in this cleft and could indicate a phosphate binding site, although not necessarily the ATP binding site since both substrates of AIR synthetase contain a phosphate group (Li, Kappock et al. 1999). Mutagenesis and affinity labeling studies of this enzyme suggest that the ATP binding site is located close to the N-terminus of the enzyme (Mueller, Oh et al. 1999). Thus, it appears that nucleotide binding is accomplished predominantly by the N-terminal domain of one subunit in the dimer, while the C-terminal domain of the opposing subunit binds the second substrate. A similar situation might be seen in the kinase members of this family. No substrate/product complex structures are available for any members of this group, although ATP has been modeled into the putative active site of one homolog (formylglycinamide ribonucleotide amidotransferase) (Anand, Hoskins et al. 2004).

Figure 2.10: The AIR synthetase-like Fold Kinase Family

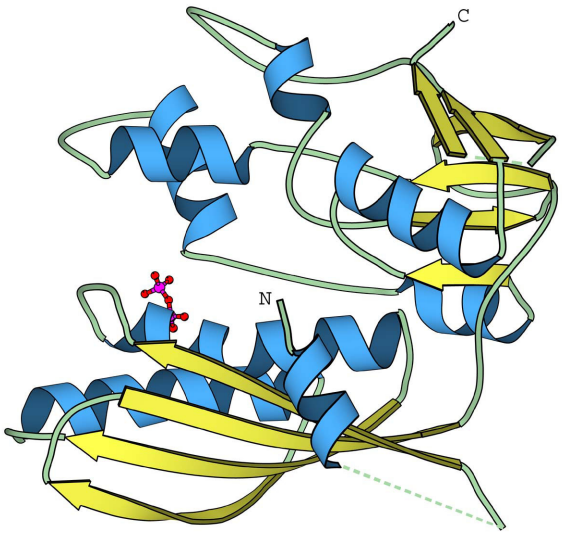


Figure 2.10: Thiamine-phosphate kianse (pdb|1vqv (Eswaramoorthy and Swaminathan *To be published*)) of the AIR synthetase-like fold group. The bound phosphate groups likely indicate the location of the active site. Dashed lines indicate disordered regions in the crystal structure.

2.4.8 Group 8: riboflavin kinase

Riboflavin kinase (RFK) is an essential enzyme in the flavin cofactor biosynthetic pathway in both prokaryotes and eukaryotes (Santos, Jimenez et al. 2000; Gerdes, Scholle et al. 2002). The structure of this kinase was unavailable at the time of the initial survey (previously Group 10) but was solved by the time of the updated survey. The core of the RFK structure (Bauer, Kemter et al. 2003; Karthikeyan, Zhou et al. 2003a) contains a 6-stranded β-barrel with Greek key topology (Figure 2.11a). This is the only kinase currently known to belong to the all-β class of proteins. The bound nucleotide is situated at one end of the β-barrel between two loop regions (L1 and L2), one of which contains a short 3_{10} helix. The solvent-exposed L1 loop and 3_{10} helix form an arch, under which the ADP phosphate tail and requisite Mg^{2+} cation bind. The adenine ring is positioned by the Pro33 and Phe97 side chains, while the ADP β-phosphate interacts with the hydroxyl group of Tyr98 and the amino group of Asn36. However, the majority of contacts with the nucleotide are made by main chain amide and carbonyl groups from the two loop regions and the 3_{10} helix (shown in magenta in Figure 2.11a). The tightly bound

Mg^{2+} ion is coordinated directly to the side chains of Thr34 and Asn36, to the α- and β-phosphates of ADP, and presumably to the γ-phosphate of ATP as well (Bauer, Kemter et al. 2003; Karthikeyan, Zhou et al. 2003b). Thr34 and Asn36 are part of the unique signature PTAN motif of riboflavin kinases which is located on a short β-strand following loop L1 and the 3_{10} helix. Drastic conformational changes induced by binding of either nucleotide or flavin ligand have been demonstrated (Bauer, Kemter et al. 2003; Karthikeyan, Zhou et al. 2003b). The mechanism of the phosphotransfer reaction in RFK appears to be direct in-line transfer of the γ-phosphate of ATP to the 5' hydroxyl group of riboflavin, which may be activated by a glutamate residue (Glu86) (Bauer, Kemter et al. 2003; Karthikeyan, Zhou et al. 2003b). The unique features of RFK appear to be that the phosphate is transferred through a hole beneath the highly dynamic Loop L1, and the proper positioning of the catalytic residues depends on binding of the substrates.

Figure 2.11: The Riboflavin Kinase and Dihydroxyacetone Kinase Families

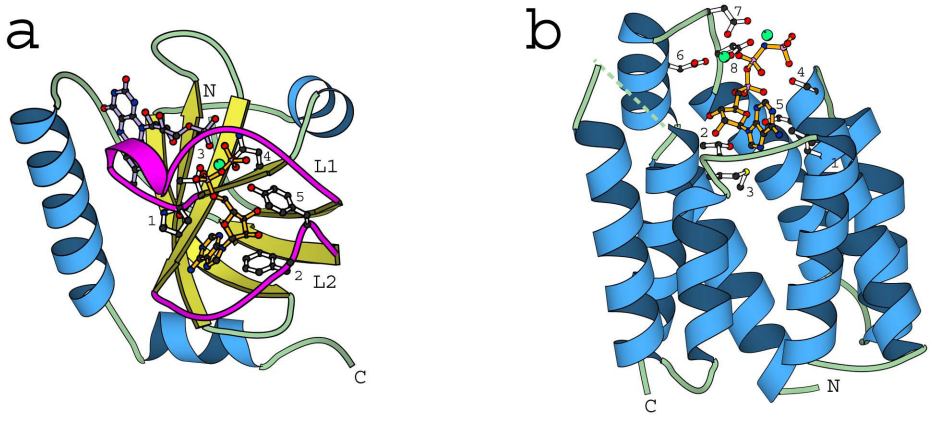


Figure 2.11: a) Riboflavin kinase (pdb|1q9s (Karthikeyan, Zhou et al. 2003b)). Loops L1 and L2 are shown in magenta. Residues **1** (Pro33) and **2** (Phe97) interact with the adenine ring of the nucleotide. Residues **3** (Thr34) and **4** (Asn36) coordinate the Mg^{2+} cation. Residues **4** (Asn36) and **5** (Tyr98) interact with the phosphate tail of the nucleotide b) Dihydroxyacetone kinase nucleotide-binding domain (pdb|1un9 (Siebold, Arnold et al. 2003)). Residues **1** (Leu435), **2** (Thr476), and **3** (Met477) pack around the adenine ring of the nucleotide. Residues **4** (Ser431) and **5** (Ser432) interact with the phosphate tail of the nucleotide. Residues **6** (Asp380), **7** (Asp385), and **8** (Asp387) are involved in coordinating the two Mg^{2+} cations. Dashed lines indicate disordered regions. In these structures, the nucleotide is colored orange, substrate molecules are purple, and Mg^{2+} cations are green balls.

2.4.9 Group 9: dihydroxyacetone kinase

Although a structural representative of this family (previously Group 17) was not available at the time of the initial kinase survey, the structure of the ATP-dependent dihydroxyacetone kinase from *Citrobacter freundii* was recently revealed (Siebold, Arnold et al. 2003). Dihydroxyacetone kinase sequences are widely distributed in organisms in all three kingdoms of life. This protein contains two regions separated by an extended linker. The N-terminal region (termed K-domain) is homologous to the non-ATP dependent DhaK protein in *E. coli* and other gram-negative bacteria. It consists of two α/β domains and is responsible for dihydroxyacetone binding. The C-terminal region (termed L-domain, homologous to the DhaL protein in *E. coli*) is the nucleotide-binding domain and is comprised of 8 antiparallel α-helices that form a closed barrel (Figure 2.11b). The α-helices are all slightly tilted away from the axis of the barrel, forming a pocket in which a phospholipid is bound. The bound ATP analog is found to be located at the top of the α-barrel. The N-terminus of one helix (H4) is pointed toward the γ-phosphate ATP and together with a glycine-rich loop between helices H3 and H4 form the primary binding site for ATP phosphates. Ser432 interacts with the ATP α-phosphate, while Ser431 interacts with the ATP β- and γ-phosphates. Two Mg^{2+} ions are coordinated by all three phosphates of ATP, and by the three highly conserved aspartates (Asp380, Asp385 and Asp387) located on a loop between helices H1 and H2. Additionally, the adenine ring is packed against several hydrophobic side chains (Leu435, Thr476, and Met477). The mechanism of phosphotransfer in DhaK is not clear since the complex conformation of the crystal structure is influenced by the crystal packing and appears not in its active form. A reaction mechanism involving a phospho-enzyme intermediate cannot be ruled out at this point. Dihydroxyacetone kinase is the only kinase known to have an all-α nucleotide-binding domain. It represents another new fold group (now Group 10) in the kinase classification scheme as its fold is unlike any other kinase with known structure.

2.4.10 Group 10: putative glycerate kinase

Putative glycerate kinase from *Thermotoga maritima* (Joint Center for Structural Genomics *To be published*) is a member of a family of proteins (previously Group 16) from eukaryotes and archaea in addition to several different bacterial species. It should be noted that this family is distinct from the glycerate kinases of Family 2i.

The fold of *T. maritima* putative glycerate kinase consists of two non-similar α/β domains (Figure 2.6b). The N-terminal α/β domain has Rossmann-like topology with the central 6-stranded β-sheet in the order of 654123. The C-terminal domain contains a 6-stranded mixed β-sheet with strand order 126345 and several helices packed on both sides of the β-sheet. The active site is likely to be in the cleft between the two α/β domains. In this structure, six highly conserved polar or charged residues are found with side chains pointing into the presumed active site (Figure 2.6b). The C-terminal domain contributes four of these highly conserved residues, while the Rossmann-like domain contributes the remaining two residues in addition to a glycine-rich loop. Each of these domains contains one highly conserved basic residue that could potentially interact with the triphosphate tail of the bound ATP: Lys47 in the Rossmann-like domain and Arg325 in the C-terminal domain. Based on the available information, it is not possible to confidently predict which domain is responsible for nucleotide binding in this putative glycerate kinase. Therefore, this family is kept as a separate fold group until its active site is characterized. The annotation for these putative glycerate kinases is based on a gene found in a 5-kb fragment that is apparently responsible for complementation in *Methylbacterium extorquens* AM1 mutants lacking glycerate kinase activity (Chistoserdova and Lidstrom 1997). However, other family members are annotated as putative glycerate dehydrogenases/hydroxypyruvate reductases based genetic analysis of the tartrate utilization pathway in *Agrobacterium vitis* (Crouzet and Otten 1995). Glycerate kinase and glycerate dehydrogenase/hydroxypyruvate reductase catalyze successive steps in the serine metabolism pathway. Therefore, the exact biochemical function of this enzyme family remains to be resolved.

2.4.11 Group 11: polyphosphate kinase

Polyphosphate kinase (PPK) synthesizes inorganic polyphosphate (polyP) by catalyzing the transfer of the γ-phosphate of ATP to a linear polymer of tens or hundreds of orthophosphate residues. Additionally, this enzyme can catalyze the transfer of a phosphate group from polyP to ADP or GDP and can generate ppppG by transferring a pyrophosphate group from polyP to GDP as well (Kuroda and Kornberg 1997). PPK sequences have been identified in many bacterial species as well as in some archaea.

Fold predictions were performed for PPK before the structure of this kinase was solved. These predictions indicated that PPK was potentially comprised of three domains: two phospholipase D-like subdomains, which are clearly recognizable by standard sequence comparison methods such as RPS-BLAST, and an N-terminal Rossmann-like domain. The phospholipase D (PLD) fold consists of a 7-stranded mixed β-sheet (strand order 1765234) flanked by several α-helices. PPK contains two PLD-like subdomains, presumably arranged in the same manner as other PLD superfamily members such as tyrosyl-DNA phosphodiesterase (pdb|1jy1 (Davies, Interthal et al. 2002)) and the homodimer of bacterial endonuclease Nuc from *Salmonella typhimurium* (pdb|1bys (Stuckey and Dixon 1999)), which is the closest structural template identified by 3D-Jury. The active site of enzymes in the PLD superfamily is located between two PLD-like subdomains. This active site is highly symmetrical due to the five equivalent highly conserved residues that are contributed by each PLD-like subdomain. Among these are two histidine residues (one from each subdomain) proposed to perform critical catalytic roles. One histidine may act as a nucleophile by attacking the phosphodiester bond that is cleaved by known PLD-like enzymes and form a phospho-histidine intermediate, while the second conserved histidine could serve as a general acid by protonating the leaving group (Stuckey and Dixon 1999). The conservation of these critical active site residues is shown in Figure 2.12. Consistent with the proposed mechanism for PLD-like enzymes, PPK has been shown to form a phospho-histidine

Figure 2.12: 2nd and 3rd Domains of PPK are Homologous to Phospholipase D

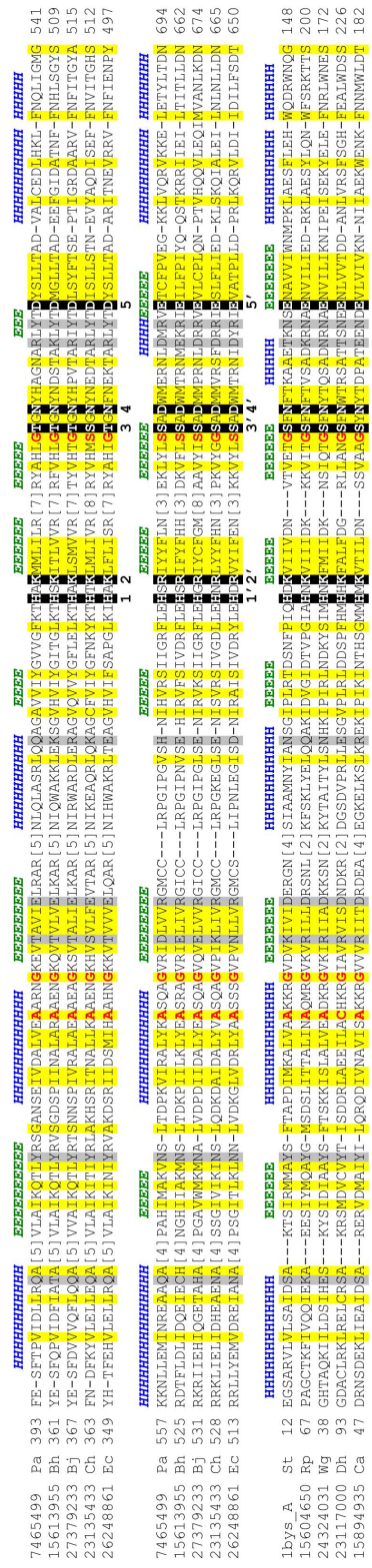


Figure 2.12: Second and Third Domains of PPK are Homologous to Phospholipase D. Alignment for representative sequences of polyphosphate kinase (PPK, gi|7465499, group 11) and phospholipase D family (pdb|1bys). Highly conserved active site residues are highlighted in black and shown in white bold text. Sequences are labeled according to the NCBI gi number or PDB code and an abbreviation of the species name. First and last residue numbers are indicated before and after each sequence. Numbers of excluded residues are specified in square brackets. Residue conservation is denoted with the following scheme: mostly hydrophobic positions, highlighted yellow; mostly charged/polar positions, highlighted grey; small residues, red bold text. Locations of predicted (gi|7465499) and observed (pdb|1bys) secondary structure elements (E, β-strand; H, α-helix) are marked above the sequences in italics and normal font, respectively. Abbreviations of species names are as follows: Bh *Bacillus halodurans*, Bj *Bradyrhizobium japonicum*, Ca *Clostridium acetobutylicum*, Ch *Cytophaga hutchinsonii*, Dh *Desulfitobacterium hafniense*, Ec *Escherichia coli*, Pa *Pseudomonas aeruginosa*, Rp *Rickettsia prowazekii*, St *Salmonella typhimurium*, Wg *Wigglesworthia glossinidia brevipalpis*.

intermediate during the phosphotransfer reaction (Ahn and Kornberg 1990; Kumble, Ahn et al. 1996) (residue **1** in Figure 2.12).

The solved structure of PPK from *E. coli* confirms that this kinase does in fact contain two PLD-like subdomains (Figure 2.13). The nucleotide is bound in the cleft between the PLD-like domains and an N-terminal α-helical bundle that was not included in the fold prediction. The N-terminal region that was tentatively suggested to be a Rossmann-like domain instead adopts an α/β fold including a mixed β-sheet with strand order 654231. Because the nucleotide binding function of PPK is carried out by the PLD-like domains, PPK denotes a separate fold group in the kinase classification since the PLD-like topology is unlike any other exiting kinase fold group.

Figure 2.13: Polyphosphate Kinase

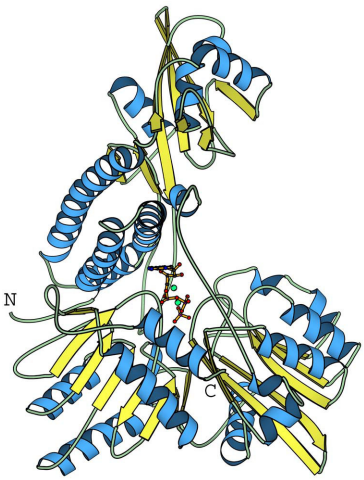


Figure 2.13: Polyphosphate kinase (pdb|1xdp (Zhu, Huang et al. 2005)) binds the nucleotide between two PLD-like domains and a 3-helix bundle. The nucleotide is orange and Mg^{2+} cations are green balls.

2.4.12 Group 12: integral membrane kinases

There are currently two known integral membrane kinases. Dolichol kinase is a member of a family of integral membrane protein cytidylyltransferases (PF01148). Dolichol kinase catalyzes the terminal step in the biosynthesis of dolichyl monophosphate (Dol-P). Similarly, undecaprenol kinase phosphorylates undecaprenol to

form undecaprenyl monophosphate (Undec-P). Undec-P is an important glycosyl carrier lipid in the cytoplasmic membrane of bacteria, and Dol-P plays a similar role in the assembly of various glycoconjugates in the endoplasmic reticulum of yeast and mammalian cells (Schenk, Fernandez et al. 2001). Thus, these two kinases have similar biological roles in that both are involved in the biosynthesis of essential polyisoprenoid glycosyl carrier lipids.

Structures are not currently available for dolichol kinase and undecaprenol kinase, although secondary structure predictions suggest that both are composed almost entire of α-helices. This is not unexpected because these are both integral membrane proteins.

2.4.13 Putative tagatose 6-phosphate kinase is removed from the classification

The putative tagatose 6-phosphate kinase (T6P kinase, previously Group 13) activity was initially suggested for the *agaZ* gene product based on the computational analysis and reconstruction of the putative N-acetylgalactosamine metabolic pathway in *E. coli* (Reizer, Ramseier et al. 1996). However, no tagatose 6-phosphate kinase activity for either AgaZ or its homolog GatZ can be detected experimentally, and genetic studies have suggested that *gatZ* is associated with tagatose-1,6-bisphosphate aldolase activity (Nobelmann and Lengeler 1996; Brinkkotter, Kloss et al. 2000). Results of transitive PSI-BLAST searches and fold predictions with 3D-Jury also suggest similarity between AgaZ and tagatose- and fructose-bisphosphate aldolases with TIM β/α-barrel fold. Alignment of the putative T6P kinases with the aldolase families revealed that several residues in the aldolases that are involved in substrate binding and catalysis are also conserved in the AgaZ/GatZ protein family (Figure 2.14). These include two histidine residues involved in the coordination of the catalytic Zn^{2+} and the aspartate residue that is proposed to protonate the substrate during the aldolase reaction (Hall, Bond et al. 2002). Thus, based on both functional study and structural prediction, it is likely that these proteins carry out an aldolase reaction rather than a kinase reaction. Therefore, this family is removed in the updated kinase classification as well.

Figure 2.14: "Tagatose 6-phosphate Kinase" is an Aldolase

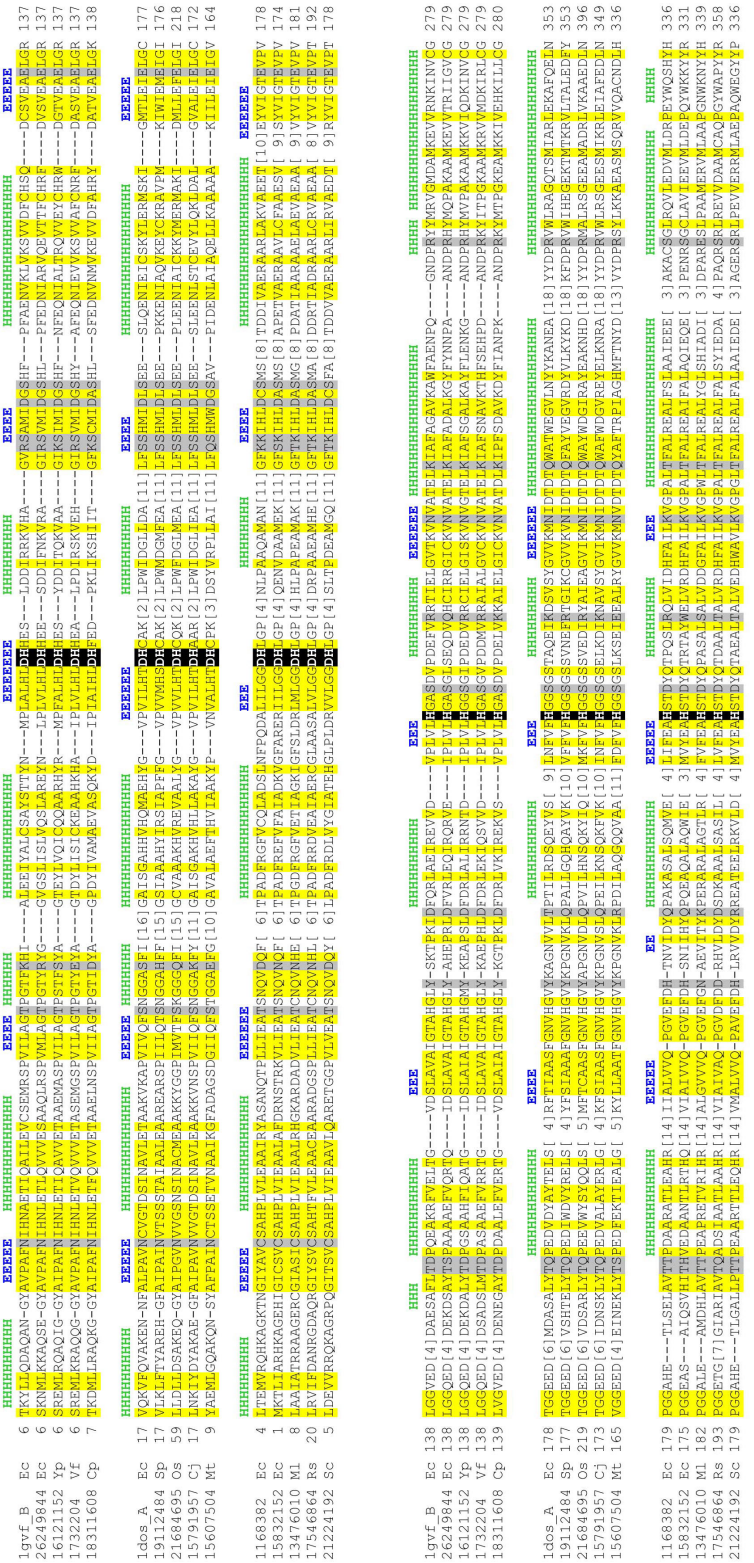


Figure 2.14: "Tagatose 6-phosphate Kinase" is an Aldolase. Alignment for representative sequences of tagatose 1,6-bisphosphate aldolase (pdb|1gvf (Hall, Bond et al. 2002)), fructose 1,6-bisphosphate aldolase (pdb|1dos (Blom, Tetreault et al. 1996)), and "T6P kinase" (gi|1168382). Highly conserved functional residues are highlighted in black and shown in white bold text. Sequences are labeled according to the NCBI gi number or PDB code and an abbreviation of the species name. First and last residue numbers are indicated before and after each sequence. Numbers of excluded residues are specified in square brackets. Residue conservation is denoted with the following scheme: mostly hydrophobic positions, highlighted yellow; mostly charged/polar positions, highlighted grey. Locations of predicted (gi|1168382) and observed (pdb|1gvf, pdb|1dos) secondary structure elements (E, β-strand; H, α-helix) are marked above the sequences in italics and normal font, respectively. Abbreviations of species names are as follows: Cj *Campylobacter jejuni*, Cp *Clostridium perfringens*, Ec *Escherichia coli*, Ml *Mesorhizobium loti*, Mt *Mycobacterium tuberculosis*, Os *Odontella sinensis*, Rs *Ralstonia solanacearum*, Sc *Streptomyces coelicolor*, Sp *Schizosaccharomyces pombe*, Vf *Vibrio furnissii*, Yp *Yersinia pestis*.

2.5 DISCUSSION

2.5.1 Common structural mechanisms shared among kinases

Although all kinases catalyze a similar phosphoryl transfer reaction, they adopt a wide variety of structural folds. The classification system presented in this work combined with a wealth of kinase structural and biochemical data is used to address the question of how these different structural folds accomplish the same biochemistry. The common structural features that influence phosphoryl transfer by kinases have been reviewed (Matte, Tari et al. 1998). Briefly, all phosphotransfer reactions contain the following three principal components: binding and orienting the phosphate donor (ATP), binding and orienting the phosphate acceptor (substrate), and catalysis of the chemical reaction.

Nucleotide binding

Several distinct modes of nucleotide binding have emerged. One recurring theme is that the nucleotide binds at the C-terminal end of β-strands and N-terminus of α-helices. This is observed in all Rossmann-like fold families and in the GHMP family. In 3 of these families, P-loop kinase, PEPCK, and GHMP, the connecting loop between the β-strand and α-helix are extended, forming the so-called phosphate-binding loop (P-loop) that wraps around the triphosphate tail of the bound nucleotide. The glycine rich nature of these loops enables them to adopt conformations such that several main chain amides are all pointed towards the bound nucleotide. Together with the positive dipole of the α-helix and some positively charged lysine or arginine side chains, a strong anion hole is created for the binding of the nucleotide.

Glycine-rich phosphate-binding loops are also observed in families of the protein kinase-like fold and ribonuclease H-like fold groups. However, in these cases, the loops are mostly β-hairpin loops (i.e. loops connecting two antiparallel β-strands). No major

contributions from helix dipoles are involved in nucleotide binding in these kinase families. However, several β-hairpin loops, such as in the case of hexokinase (ribonuclease H-like), may congregate at the active site with their main chain amides interacting with the phosphate. In the above two distinct nucleotide-binding modes, the protein main chain interacting with the nucleotide is prominent. This is probably one of the reasons why the ATP binds at roughly the same place in all Rossmann-like families.

Nucleotide binding in the ferredoxin-like group and TIM-barrel group are completely different. No commonly shared local structural motifs are observed across families. In general, kinases in these two fold groups mainly use positively charged protein side chains to interact with the nucleotide phosphates. As discussed before, nucleotide binding differs greatly between the families of the ferredoxin-like group, probably as a result of a lack of interactions with protein backbones.

Binding of phosphoryl acceptor substrates

Binding and orientation of the phosphate-accepting substrate in kinases depends on interactions between the substrate and strategically placed active site residues. The details of such interactions are, of course, contingent upon the specific activity of the kinase in question. Extra structural motifs or domains in addition to the nucleotide-binding core are usually necessary for the recognition of the substrate. Since the size and structure of kinase substrates varies drastically from a small molecule of a few atoms to a whole protein, the substrate binding motifs also vary significantly. One common phenomenon associated with the substrate binding is induced conformational changes. In some cases, such as pyruvate phosphate dikinase and phosphoglycerate kinase, drastic domain movements occur in order to bring the substrate to the active site (Herzberg, Chen et al. 1996; Bernstein, Michels et al. 1997). While in other cases, such as protein kinase and GHMP kinase family members, local conformational changes, such as closing a loop conformation, is sufficient to sequester the substrate and ensure the transfer of the phosphoryl group from ATP to the substrate.

Metal binding and catalysis

Almost all kinases require a divalent metal cation in order to function. Mg^{2+} usually activates ATP for catalysis by weakening the bond between the β- and γ-phosphates and assists in properly orientating the γ-phosphate for phosphotransfer. Metal cations are usually coordinated by a conserved glutamate, aspartate, or other hydroxyl-containing residue in the active site. In some cases, however, the Mg^{2+} cations are positioned by coordinated water molecules and have no direct liganding to the enzyme. Several kinases utilize additional metal cofactors such as a secondary Mg^{2+} cation, a Mn^{2+} cation, or a K^{+} cation (Tari, Matte et al. 1997; Larsen, Benning et al. 1998; Machius, Chuang et al. 2001).

There are three main catalytic mechanisms that are widely used by kinases: transition state stabilization, acid-base catalysis, and ping-pong (or double displacement) mechanisms. In general, there is little correlation between the structural fold group and the mechanism of catalysis utilized. One example of different mechanisms in the same fold group includes the ATP-grasp family and the protein kinase family of Group 1. The mechanism of pyruvate phosphate dikinase of the ATP-grasp family involves the reversible phosphorylation of a histdine residue (Spronk, Yoshida et al. 1976; Goss, Evans et al. 1980). Although the mechanism utilized by protein kinases is currently a matter of debate, the phosphotransfer reaction in this family is thought to proceed either via a proposed simultaneous transfer mechanism (Hart, Hillier et al. 1998), which falls in the category of transition state stabilization, or via acid-base catalysis. A second example includes the families of the ferredoxin-like fold group. Nucleoside-diphosphate kinase (NDP) kinase and histidine kinase both utilize ping-pong mechanisms. Although the possibility of acid-base catalysis cannot be eliminated in arginine kinase of the guanido kinase family, the primary factor in the mechanism of this enzyme appears to be transition state stabilization and precise substrate orientation in the active site (Zhou, Somasundaram et al. 1998).

Within most kinase families, all enzymes usually utilize the same type of catalytic mechanism. All protein kinases, for example, are thought to have similarly catalyzed

reactions. However, there are exceptions to this tendency. In the GHMP family, for example, the homoserine kinase mechanism is proposed to proceed via a simultaneous transfer reaction (transition state stabilization) (Krishna, Zhou et al. 2001), while the archaeal shikimate kinase reaction may follow a ping-pong mechanism (Daugherty, Vonstein et al. 2001). In a second example, reactions catalyzed by P-loop family kinases may proceed via different mechanisms. Acid-base catalysis is the probable mechanism of P-loop enzymes such as phosphoribulokinase and shikimate kinase for the production of phosphorylated metabolites (Miziorko 2000), while UMP/CMP kinase of the same family has been suggested to utilize a simultaneous transfer mechanism similar to the one proposed for protein kinases (transition state stabilization) (Hutter and Helms 2000).

2.5.2 Distribution of kinases in genomes

The distribution of kinases from the genomes of several representative species is shown in Table 2.9. The protein kinase (PK) family of Group 1 is by far the largest family in terms of number of sequences from the non-redundant database. The PK family contains over one-half of the sequences in the initial survey and over one-third of the sequences in the kinase survey. In the selected representatives of eukaryotic species (*Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae*), sequences belonging to the PK family account for between 65% and 85% of all identified kinases in these genomes. In prokaryotes, however, the two-component system (histidine kinase) is the predominant method of signal transduction. This is reflected by the small number of proteins in the PK family from the genomes of *Escherichia coli* and *Methanococcus jannaschii*. The non-signaling kinase groups, which include all families except the protein kinase family and the histidine kinase family, predominantly contain metabolic kinases. For these families, the distribution of kinase sequences is approximately equal for five of the six representative species (Table 2.9). Although the percentage of metabolic kinases differs between the organisms, the actual numbers of these enzymes in each of the representative genomes are rather similar,

ranging from 59 in *S. cerevisiae* to 100 in human. Notably, there are only 31 identified kinase genes in *M. jannaschii*. This may be due to the smaller genome size of *M. jannaschii*, as well as unique aspects of archaeal metabolism. It should also be pointed out these totals are not the true number of kinases in these genomes. Although these representative genomes are completely sequenced, there are still many "hypothetical" (unannotated) genes within them. Furthermore, genes coding for ~65 known kinase activities are not yet identified in any organisms.

Table 2.9: Distribution of Kinases in Representative Genomes

	Number of Kinase Sequences		All kinases	% of all identified proteins in genome
	Protein kinase family	Non-signaling kinase families		
Homo sapiens	391	100	492	1.5%
Drosophila melanogaster	273	91	365	2.5%
Caenorhabditis elegans	473	76	549	2.7%
Saccharomyces cerevisiae	126	59	187	3.0%
Escherichia coli	1	76	107	2.5%
Methanococcus jannaschii	2	29	31	1.8%

Table 2.9: Distribution of Kinases in Representative Genomes. The non-signaling kinase families include all families except the protein kinase family and the histidine kinase family.

2.5.3 Convergent evolution of kinase activities

There are several cases of the same kinase activity that exists in unrelated fold families, reflecting convergent evolution of the same function. For example, Galperin *et al*. have identified analogous enzymes in fructokinase, 6-phosphofructokinase, and gluconokinase (Galperin, Walker et al. 1998). Comparisons of homoserine kinase, phosphomevalonate kinase, shikimate kinase, glucokinase, 1-phosphofructokinase, uridylate kinase, and pantothenate kinase from different folds are summarized below.

Homoserine kinase is found in two distinct fold groups: the protein kinase family and the GHMP kinase family. The protein kinase family contains homoserine kinases from several proteobacterial species (eubacteria). The GHMP kinase family includes homoserine kinases from the majority of eubacteria, from archaebacteria, and from eukaryotes. It is interesting to note that the same mechanism of catalysis has been proposed for both the protein kinase family and the GHMP family (the "synchronous shift" mechanism). One surprising feature is homoserine kinase's use of the protein kinase fold, which is generally utilized for the accommodation of very large substrates.

Phosphomevalonate kinase (PMK) and s*hikimate kinase* (SK) are each found in both the P-loop kinase family and the GHMP kinase family (Smit and Mushegian 2000; Daugherty, Vonstein et al. 2001). PMKs from higher eukaryotes (including human, pig, and fruit fly) belong to the P-loop kinase family. All other identified PMK sequences (such as those of *Saccharomyces cerevisiae*, *Staphylococcus aureus*, and *Streptococcus pneumoniae*) are found in the GHMP kinase family. SK also contributes members to each of these two families. The P-loop kinase family includes primarily eubacterial SKs while the GHMP kinase family contains archaebacterial SKs (Daugherty, Vonstein et al. 2001). Currently, the only PMK structure available is from the GHMP kinase family, and the only SK structures available are from P-loop kinase family.

Glucokinase is found in both the ribokinase-like family of the Rossmann-like fold group and the ribonuclease H-like fold group. Glucokinase from archaeal species such as *Thermococcus litoralis* and *Pyrococcus furiosus* belong to the ribokinase-like family. The ribonuclease H-like family contains glucokinases from eukaryote, eubacterial, and a few archaebacterial species. Unlike the ATP-dependent ribonuclease H-like glucokinases, the archaeal glucokinases of the ribokinase-like family are ADP-dependent. The modes of nucleotide binding differ between these two families.

1-phosphofructokinase can be found in both the phosphofructokinase-like and ribokinase-like families. The sequences found in the ribokinase-like family are from bacterial species, while only human sequences with this activity are found in the phosphofructokinase-like family. Both of these families belong to the Rossmann-like

group, and both families are believed to utilize acid-base catalysis in their reactions. Thus, the core of the folds and the mechanisms of phosphotransfer are expected to be similar for 1-phosphofructokinases from each family. However, because the nucleotide-binding patterns are somewhat different between these two families, the precise location and orientation of the bound ATP will most likely differ between human and bacterial 1-phosphofructokinase.

Uridylate kinase is also found in two different Rossmann-like families. Uridylate kinase from *Leishmania major* and *Saccharomyces cerevisiae* belong to the P-loop kinase family. Uridylate kinase sequences in the aspartokinase family are predominantly from bacterial and archaebacterial species. Again, since both families are in the Rossmann-like group, the core of the uridylate kinase structures will be similar while specificities of nucleotide binding will differ between the uridylate kinase representatives in the two families.

Pantothenate kinase provides another interesting example. Prokaryotic pantothenate kinase is known to belong to the P-loop kinase family (Yun, Park et al. 2000). However, eukaryotic pantothenate kinase is predicted to adopt ribonuclease H-like fold. Thus, the prokaryotic and eukaryotic versions of this enzyme are likely to differ in fold, in mode of nucleotide binding, and in catalytic mechanism.

The examples above describe cases in which nature has developed the same activity in multiple ways. The opposite situation, in which the same structural fold is used for many different substrate specificities, is readily observable as well. Examples of families in which one structural fold accounts for kinase activity on many different substrates include the P-loop kinase family, the ribokinase-like family, and the ribonuclease H-like family.

2.5.4 Correlation of structural fold with placement in cellular pathway

One of the few generalities that can be made concerning the correlation between structural fold and cellular pathway is that the protein kinase fold is dedicated

predominantly to cellular signaling. The vast majority of the members of the protein kinase family participate in signal transduction. Furthermore, the Rossmann-like fold in kinases is apparently utilized exclusively in metabolic pathways. However, the types of metabolic pathways that the kinases participate in vary between each of the families of the Rossmann-like group. For example, most kinases in the ribokinase-like family are involved in carbohydrate metabolism, although a few do participate in the metabolism of nucleotides or vitamins and cofactors. The aspartokinase family, however, has many members that participate in amino acid metabolism in addition to a few that are involved in energy metabolism or nucleotide metabolism. P-loop kinases represent the Rossmann-like family whose members participate in the widest variety of metabolic pathways types. While the largest fraction of P-loop kinase family members participate in nucleotide metabolism, a substantial number also function in the metabolism of lipids, carbohydrates, amino acids, and other types of molecules. As a whole, Group 2 (Rossmann-like) kinases are involved in the entire scope of metabolic pathway types, including carbohydrate, lipid, amino acid, nucleotide, cofactor, vitamin, and energy metabolism.

Kinases of the ferredoxin-like fold group also have evident partialities in terms of pathway types. Members of the guanido kinases family function solely in amino acid metabolism, and histidine kinases are signaling enzymes. The other two families in this group each contain only one kinase member: nucleoside-diphosphate kinase functions in nucleotide metabolism while HPPK participates in vitamin metabolism pathways.

Although GHMP kinases participate in several different metabolic pathways, such as carbohydrate, amino acid, and lipid metabolism, their role in the isoprenoid biosynthesis pathways is particularly prominent. Notably, members of the GHMP kinase superfamily (mevalonate kinase, phosphomevalonate kinase, and mevalonate pyrophosphate decarboxylase) catalyze three consecutive steps in the early mevalonate pathway. One other GHMP kinase, 4-(cytidine 5'-diphospho)-2-*C*-methyl-D-erythritol kinase, participates in the recently characterized non-mevalonate isoprenoid biosynthesis pathway (Luttgen, Rohdich et al. 2000). The essentiality of these enzymes have

identified them as potential anti-bacterial drug targets (Jomaa, Wiesner et al. 1999; Lichtenthaler, Zeidler et al. 2000; Wilding, Brown et al. 2000).

Trends in other groups are less evident. Members of the ribonuclease H-like group participate mainly in carbohydrate metabolism, although a significant number of kinases are involved in other types of metabolic pathways as well. Notably, ATP-grasp and TIM β/α-barrel, two of the most widespread protein folds, are adopted for only 3 and 1 kinase activities, respectively. However, each family includes participants in 3 different types of metabolic pathways (lipid, carbohydrate, and energy metabolism for ATP-grasp and nucleotide, carbohydrate, and energy metabolism for TIM β/α-barrel kinases). This may reflect the overall functional diversity of both folds.

2.5.5 Comprehensive structural annotation of kinases

Of the 25 kinase families, 22 currently have at least one homolog with a solved structure (Tables 2.5-2.7). The structural folds of each domain within one additional family (polyphosphate kinase) are predicted as discussed in section 2.4.11. The two remaining families are integral membrane kinases. Although the tertiary structure of dolichol kinase and undecaprenol kinase are not yet determined, secondary structure predictions indicate that both of these families adopt all α-helical conformations. Thus, structural annotations of all biochemically characterized kinase families are now revealed, including fold descriptions for all globular kinases, and the kinase fold groups listed in Tables 2.5-2.7 present the complete structural depiction of this entire functional class of proteins. The structural folds adopted by kinases include some of the most widely spread protein folds, including the Rossmann-like fold, ferredoxin-like fold, and TIM β/α-barrel fold. The kinase fold repertoire also includes representatives of all major classes (all-α, all-β, α+β, α/β) of protein structures, demonstrating that nature has found ways to utilize all varieties of secondary structure combinations to carry out the kinase reaction.

2.6 CONCLUSIONS

A comprehensive survey of all available kinase sequences and structures has been performed. All available kinases (~59,000 sequences and 700 structures) have been classified into 25 distinct families of homologous proteins and 12 fold groups reflecting structural similarity. All kinase families, with the exception of the integral membrane kinases, are now associated with a known or predicted structural fold. Therefore, the kinases are the first large functional class of proteins with a comprehensive structural annotation for its known members. This work presents the final global picture of this entire class of enzymes, which are now known to adopt folds from all major structural classes (all-α, all-β, α+β, α/β).

Additionally, the robustness of this classification was demonstrated by the completion of a comprehensive update. This updated survey showed that despite a 3-fold increase in the number of kinase sequences and 2-fold increase in the number of kinase structures, the framework of the initial classification remains sufficient for describing all available kinases. This update also revealed that no fold predictions made in the initial kinase survey were shown to be incorrect.

The completion of this classification allowed for the investigation of how different structural folds carry out the same fundamental aspects of the kinase phosphotransfer reaction. The classification also revealed cases of convergent evolution of identical biochemical activities from unrelated protein families, and many examples of enzymes that have diverged from the same protein ancestor to accomplish different specific kinase activities.

CHAPTER 3:
Structural Classification of Small Disulfide-Rich Protein Domains

3.1 INTRODUCTION

3.1.1 Background

The structures of very small proteins often lack an extensive hydrophobic core and possess secondary structure elements that are small and irregular. These proteins are generally stabilized either by binding a metal ion (most commonly, zinc (Krishna, Majumdar et al. 2003)) or by the formation of disulfide bonds. Disulfide bonds have traditionally been presumed to stabilize protein structures by reducing the conformational freedom of the protein in the unfolded state, therefore reducing the entropy of the unfolded state relative to the folded state (Flory 1956; Anfinsen and Scheraga 1975; Thornton 1981). Another theory proposes that the stabilizing influence of these cross-links is enthalpic, whereby the presence of the disulfide bonds destabilize the denatured form of the protein by sterically inhibiting certain potential hydrogen bonding groups from forming satisfied donor-acceptor pairs (Doig and Williams 1991). It has also been suggested that both entropic and enthalpic effects contribute to the stabilizing capacity of disulfide bonds (Betz 1993). Although these cross-links are, in most cases, mainly responsible for maintaining the proper fold of the protein and are therefore only indirectly essential for protein function, there are examples in which reduction or oxidation of these bonds alters protein activity (Aslund and Beckwith 1999; Yano, Kuroda et al. 2002).

Small protein domains in which disulfide bonds form the scaffold of the protein are often referred to as disulfide-rich. In this study, a typical disulfide-rich domain is described by the following characteristics: small in size (usually <100 residues), lacking an extensive hydrophobic core, having few secondary structure elements, and fold stabilization primarily due to two or more disulfide bonds in close proximity. These proteins encompass a wide variety of functions, such as growth factors, toxins, enzyme

inhibitors, and structural or ligand-binding domains within larger polypeptides. Several classes of disulfide-rich proteins have been of interest to researchers for medical reasons, such as insulin and related growth factors or ion channel inhibiting toxins. Other disulfide-rich proteins have been the focus of folding experiments, with bovine pancreatic trypsin inhibitor being the most thoroughly studied example (Creighton 1992; Creighton 1997). These folds have also been proposed as scaffolds for drug design (Menez 1998; Craik, Simonsen et al. 2002) and mimetics of protein interacting surfaces (Vita, Drakopoulou et al. 1999).

Protein classification on the basis of structural similarity and evolutionary relatedness is a common means of organizing biological data for the purpose of studying various aspects of sequence-structure-function relationships in proteins, such as structure prediction or identification of functionally important residues. Evolutionary and structural neighbors of large (>100 residues), globular proteins can often be identified using popular sequence and structure comparison tools such as PSI-BLAST (Altschul, Madden et al. 1997) and Dali (Holm and Sander 1995). However, automatic methods generally tend to be unreliable for small proteins, due to the short length of these polypeptide chains. Classification of small protein domains is consequently a non-trivial task and one that frequently requires considerable manual analysis.

Classification schemes for disulfide-rich domains have previously been constructed using automated tools that compare the geometry and topology of disulfide bonds. The KNOT-MATCH program clusters proteins based on structural superposition of the disulfide bonds (Mas, Aloy et al. 1998; Mas, Aloy et al. 2001). Another approach classifies proteins according to their "disulfide signature", which considers disulfide connectivity and the loop lengths between cysteine residues (Gupta, Van Vlijmen et al. 2004; van Vlijmen, Gupta et al. 2004). However, the evolutionary relatedness among protein groupings identified by these approaches must be carefully interpreted, as these methods do not address established indicators of homology or biologically relevant factors, such as sequence similarity, protein function, fold topology, or other structural features beyond disulfide bonding patterns. A number of other studies have examined

specific subsets of disulfide-rich domains, focusing on a particular family (e.g. toxins from snails (Espiritu, Watkins et al. 2001) or spiders (Escoubas, Diochot et al. 2000)), structural motif (e.g. the KNOTTIN website (Gelly, Gracy et al. 2004)), or function (e.g. protease inhibitors; MEROPS (Rawlings, Tolle et al. 2004)). Although nearly all disulfide-rich domains are included in the comprehensive SCOP database (Murzin, Brenner et al. 1995), this is not a convenient tool for studying this group of proteins as a whole because the disulfide-rich domains are distributed among several structural classes (small proteins, all-α proteins, peptides, etc).

3.1.2 Objectives

In order to understand the structural and functional diversity among all available small disulfide-rich proteins, a comprehensive classification of these domains has been performed. Due to the nature of small protein domains, construction of this classification was based predominantly on manual sequence and structure analysis. The hierarchy of this classification is comprised of two levels, such that the disulfide-rich domains are evaluated in terms of both their structural and evolutionary relatedness. Based on this survey, the variety of structural folds adopted by disulfide-rich domains are examined, and the distant homology between previously unlinked domains is described. Disulfide bonding patterns of these domains are evaluated, and examples of convergent and divergent evolution of functions performed by these proteins are identified.

3.2 METHODS

3.2.1 Identification of disulfide-rich protein domains

A protein is considered to potentially be disulfide-rich if the structure contains 2 disulfide bonds within 23 Å. This distance cutoff was determined empirically based on

protein domains previously noted as disulfide-rich in the "Small proteins" class of the SCOP database. A locally mirrored version of the PDB (current through August 2, 2005) was searched for structures containing 2 or more disulfide bonds within 23 Å. A disulfide bond was assumed to exist between two cysteine residues if their gamma sulfur atoms were less than 3.5 Å apart. Sequences of individual chains from PDB structures identified by this automated search were extracted and clustered on the basis of sequence identity using the BLASTCLUST program (I. Dondoshansky and Y. Wolf, unpublished; ftp://ftp.ncbi.nih.gov/blast/) with a 50% identity threshold and length coverage threshold of 90% on each sequence. A representative of each cluster was examined in order to identify and exclude non-disulfide-rich chains within PDB structures as well as structures in which cysteine side chains contribute to metal-binding rather than disulfide bonds.

For the purposes of this study, only protein domains with structural folds stabilized primarily by the formation of disulfide bonds, rather than by the hydrophobic core of the protein, are of interest. Such proteins typically have a very small hydrophobic core and few secondary structure elements. Therefore, proteins with a significant hydrophobic core and many secondary structure elements were removed from the set of structures that were identified in the automated search. For example, the structure of *Macadamia integrifolia* antimicrobial protein MiAMP1 (pdb|1c01 (McManus, Nielsen et al. 1999)), which contains 3 disulfide bonds but also a substantial hydrophobic core (8-stranded β-sandwich with Greek key topology), was excluded from this classification because the disulfide bonds appear to be incidental to the stability of the protein's structural fold. Likewise, proteins for which non-disulfide-rich homologs or structural analogs are known were also excluded, such as the *Aspergillus giganteus* antifungal protein AGAFP (pdb|1afp (Campos-Olivas, Bruix et al. 1995)), which adopts an OB-like fold. Structures such as these are considered to be better described by their fold topology than by their disulfide bonds, and are therefore more appropriately classified with their non-disulfide-rich structural neighbors. Such cases were identified by manual examination of cluster representatives. Domains are included in this survey only if they are continuous, with the exception of circular permutations and multi-chain domains.

3.2.2 Classification of disulfide-rich protein domains

The disulfide-rich protein domains identified from the PDB were classified according to a two-tier hierarchy. The first tier is the fold group level, which is based on structural similarity between protein domains. Domains in the same fold group share a common structural core with topology that is either identical or related by circular permutation. Fold groups were determined by visual inspection. The second tier is the family level, which reflects an evolutionary relationship between domains. Homology between members of a family was established based primarily on sequence similarity, but factors such as functional similarity and conservation of key functional residues were also considered when that information was available.

The InsightII package was used to visualize and superimpose the structures of the disulfide-rich protein domains. Multiple structure-based alignments were manually constructed for each family and fold group based on the superpositions made in InsightII. These alignments are available at ftp://iole.swmed.edu/pub/disulf_aln.

3.2.3 Evaluation of disulfide bonding patterns

Disulfide bonding patterns in the set of small, disulfide-rich domains identified in this study were analyzed according to the intrinsic properties of symmetry and reducibility, as defined by Benham and Jafri (Benham and Jafri 1993). A bonding pattern has symmetry if it reads the same from N→C and C→N. For example, if a domain has 3 disulfide bonds where the first cysteine is bonded to the second, the third to the fourth, and the fifth to the sixth, then it will read as *aabbcc* from both directions. A bonding pattern is reducible if a single cut can separate it into two discrete subpatterns, where no disulfide bond is split between the subpatterns. For example, the pattern *ababcc* is reducible because it can be cut into *abab* and *cc*, but the pattern *abcabc* is irreducible because it cannot be split into self-contained subpatterns.

The number of symmetrical and reducible patterns have been tabulated in two ways: by the bonding pattern of each family (defined as the pattern of the disulfide bonds conserved in >80% of the family members) and by the bonding pattern of each representative domain after clustering all of the domains in the classification (95% identity and 95% length coverage). Both measures are considered because neither method alone provides an ideal sample: the number of families provides a very small sample size, but the counts given by representatives are biased in favor of overpopulated families such as scorpion toxins and epidermal growth factor-like domains. Furthermore, seven families from the classification with members having different disulfide bonding patterns were excluded from this analysis. For example, two domains in the cellulose binding/docking domain family (fold group 18) have bonding pattern *aabb* (pdb|1e8p, 1e8q (Raghothama, Eberhardt et al. 2001)) while the other two domains in this family have bonding pattern *abba* (pdb|1e8r, 1qld (Raghothama, Simpson et al. 2000)). For consistency, members of these seven families are also excluded in the calculations with representative domains.

The predicted fraction of reducible patterns with N disulfide bonds expected by random, equiprobable connections between cysteines can be described by the number of possible reducible patterns with N bonds divided by the total number of possible patterns with N bonds. Similar calculations give the predicted number of symmetric patterns with N disulfide bonds.

3.3 RESULTS OF DISULFIDE-RICH DOMAIN CLASSIFICATION

3.3.1 Results of disulfide-rich domain classification

Disulfide-rich domains identified in the PDB

Structures of 2945 small disulfide-rich protein domains were detected in the PDB as described in section 3.2. These domains are found in 2578 individual PDB chains

from 1596 PDB structures. However, the notable interest researchers have taken in these proteins is reflected in the high degree of redundancy within this set (due to identical chains within one PDB structure or multiple structures of the same protein). Upon clustering the sequences of these 2945 domains at 95% identity with 95% length coverage, the number of representatives is reduced to 963 domains. Although the "unique" representatives comprise only ~33% of the original set, a similar reduction is not achieved by further decreasing the identity among clusters: clustering at 50% identity with 95% coverage results in 696 disulfide-rich domains (~24% of the original set). The protein domains in this classification are an average of 57 $\pm$ 29 residues in length and contain an average of 3 $\pm$ 1 disulfide bonds. Most of these domains (>96%) are from eukaryotic organisms.

Disulfide-rich domains are classified into fold groups and families

The 2945 disulfide-rich protein domains are arranged into 41 fold groups according to structural similarity (Tables 3.1-3.3). Domains within the same fold group share a common structural core comprised of secondary structure elements found in the same spatial arrangement with topology that is either identical or related by circular permutation. Despite this structural similarity, homology between all domains within a fold group is not implied. Within each fold group, the disulfide-rich domains are classified into families based on evolutionary relationships between members, which are inferred from the similarity of protein sequences, structures, and functions. The 2945 domains in this classification are arranged into 98 families of homologs.

Most of the different families within one fold group are likely the result of convergent evolution of unrelated proteins to a similar structural fold. However, some families may contain distantly related homologs for which there is currently insufficient sequence and functional information to confidently support homology between them. For example, members of fold group 3 are structurally characterized by a disulfide-bonded 3-helix bundle with right-handed connections between the α-helices (Figure 3.1b). There are currently 6 distinct families of disulfide-rich domains within this fold group.

Table 3.1: Small Disulfide-Rich Domain Classification, Fold Groups 1-12

Fold group	*Common structural core*	*Families*	*# of members*		*Representative*
			all domains	*95% identity*	
1	small, distorted α-hairpin	κ-hefutoxin-like	4	3	1hp9, A1-A22
		immunodominant domain of attachment protein G	3	2	1brv, 171-189
		endothelin-like	8	6	1srb, 1-21
		integrin αVβ3 subdomain	4	2	1jv2, B601-B636
		cellulase subdomain	44	6	1a39, 41-74
		IgE receptor antagonist	6	2	1kcn, A1-A21
2	α-hairpin	cytochrome c oxidase, subunit VIb	14	2	1ocr, H7-H85
		cytochrome bc1 complex, non-heme 11 kDa protein ("hinge")	18	5	1bcc, H13-H78
		cytochrome c oxidase copper chaperone	1	1	1z2g, A1-A69
		enterotoxin B	1	1	1ehs, 1-48
		Ole e 6 pollen allergen	1	1	1ss3, A1-A50
		attractin	1	1	1t50, A1-A58
		neurotoxin B-IV	1	1	1vib, 1-55
		vanabin 2	1	1	1vfi, A1-A95
3	3-helix bundle, right-handed	protozoan pheromones, ER-1-like	6	5	1erp, 1-38
		anaphylotoxin C5a	3	3	1cfa, A1-A71
		P8-MTCP1	3	2	1hp8, 1-68
		Notch/ DSL/ LNR domain	1	1	1pb5, A1-A35
		sea anemone toxin K	4	3	1bgk, 1-37
		CRISP family, helical bundle subdomain	5	3	1rc9, A181-A221
4	3-helix bundle, left-handed	insulin-like	242	15	7ins, B1-B30, A1-A21
		helical subdomain of serine carboxypeptidase-like	5	2	1cpy, 181-252
		molt-inhibiting hormone	1	1	1j0t, A0-A77
5	3-helix irregular bundle with disulfide bonds to N- and C-terminal extensions	frizzled family	8	2	1ijx, D2-D123
6	4 small α-helices, non-globular array	domain II of osmotin-like family	18	4	1aun, 129-177
7	5-helix globular array I	protozoan phermone, ER-23	1	1	1ha8, A1-A51
8	5-helix globular array II	tetraspanin family ectodomain	4	1	1g8q, A113-A202
9	5-helix "hollow" array	elicitins	6	2	1bxm, -1-98
		GFRα1 domain 3	1	1	1q8d, A239-A346
10	2 antiparallel disulfide-linked α-helices and a Ca^{2+}-binding loop	phospholipase A2	271	52	1hn4, A-5-A124
11	β-hairpin	antimicrobial β-hairpin	11	10	1hvz, A1-A18
		arylsulfatase, β-hairpin subdomain	7	1	1auk, 151-177
		subdomain of Fr-MLV envelope glycoprotein receptor-binding domain	1	1	1aol, 67-95
12	3-strand β-sheet, antiparallel, strand order 123	locust serine protease inhibitors	10	8	1pmc, 1-36
		fibronectin type I module	13	8	1qgb, A61-A109
		midkine	2	2	1mkn, A1-A59
		thrombospondin type I repeat	4	4	1lsl, A416-A472
		hormone binding domain of CRF receptor	1	1	1u34, A15-A133
		β-microseminoprotein, N-terminal domain	1	1	1xhh, A1-A49

Table 3.2: Small Disulfide-Rich Domain Classification, Fold Groups 13-25

Fold group	*Common structural core*	*Families*	*# of members*		*Representative*
			all domains	*95% identity*	
13	3-strand β-sheet, antiparallel, strand order 132	mammal defensin-like/ sea anemone toxin-like	57	22	1dfn, A2-A31
		Bowman-Birk inhibitor/ bromelain inhibitor	38	25	1bbi, 25-51
		Amb V ragweed allergen	3	1	1bbg, 1-40
14	3-strand β-sheet, mixed, strand order 132	domain III of malarial parasite apical membrane antigen I	2	2	1w8k, A387-A453
		anti-HIV peptide RP 71955	2	1	1rpb, 1-21
		chordin-like cysteine-rich repeat	1	1	1u5m, A44-A73
		IGFBP family, N-terminal domain	1	1	1wqj, B3-B39
15	3-strand β-sheet, mixed, strand order 312	crambin-like *(α-hairpin inserted in β-sheet)*	24	9	1cbn, 1-46
		fungal pathogen protein NIP1	2	2	1kg1, A29-A60
16	3-"strand" bundle; 2 strands form β-hairpin	TNF receptor family repeats	147	32	1d0g, T102-T130
		vascular endothelial growth factor	6	2	1vgh, 1-27
17	2 parallel β-hairpins	domain III of osmotin-like family	18	5	1aun, 49-80
18	2 perpendicular hairpins	cellulose binding/docking domains	4	2	1e8r, A20-A69
19	5 β-strands in 2 parallel layers; each layer is antiparallel	CCP modules/ SCR domains/ Sushi domains	170	37	1g40, B65-B125
20	5 β-strands in 2 perpendicular layers; each layer is antiparallel	Kringle-like/ fibronectin type II module	99	31	1ks0, A1-A59
21	interconnected 3-"strand" subdomains	disintegrins	15	9	1fvl, 1-70
22	irregular β-sandwich	methylamine dehydrogenase, L chain	16	2	2bbk, L7-L131
23	knottin-like I: 4 cysteines forming disulfide crossover are located on 4 structure elements; elements 1-2-3 w/right-handed connection, 2-3-4 w/left-handed connection	spider toxin/ ω-conotoxin/ IGFBP knottin-like domain/ VHv1.1 viral protein subdomain/ plant enzyme inhibitor/ gurmarin/ agouti-like	111	75	1lmm, A1-A40
		scorpion toxin-like/ insect and plant defensin-like	124	82	1agt, 1-38
		Kalata-like cyclotides	14	12	1kal, 1-29
		cellulose-binding domain of cellobiohydrolase I	6	1	1az6, 1-36
		satiety factor CART, subdomain	1	1	1hy9, A49-A89
		plant lectin-like	109	19	1hev, 1-43
		colipase-like	16	7	1lpa, 42-90
		cysteine knot cytokines	125	35	1aoc, A83-A175
24	knottin-like II: 4 cysteines forming disulfide crossover are located on 4 structure elements; left-handed connections	snake toxin-like	158	44	1idg, A1-A74
		leech serine protease inhibitor-like	26	11	1c9t, J7-J59
		granulin-like repeat, N-terminal domain	5	4	1fwo, A1-A35
25	knottin-like III: first 2 of 4 cysteines forming disulfide crossover are located on 1 irregular/bulging structure element	EGF-like	286	88	1nub, A53-A78
		cysteine-rich repeats of EGF receptor family ectodomain	153	65	1s78, A171-A192
		CRISP family, knottin-like subdomain	6	3	1rc9, A166-A180
		DPY module	1	1	1oig, A1-A24
		elafin-like	3	3	2rel, 1-57
		invertebrate antimicrobial (chitin-binding) proteins	2	2	1dqc, A1-A73
		bubble protein	1	1	1uoy, A1-A64

Table 3.3: Small Disulfide-Rich Domain Classification, Fold Groups 26-41

Fold group	*Common structural core*	*Families*	*# of members*		*Representative*
			all domains	*95% identity*	
26	inverted knottin: disulfide crossover is stacked in opposite direction as fold groups 23/24/25	trefoil	20	4	1pcp, 1-52
		PSI domain	14	6	1olz, A480-A533
		myeloperoxidase subdomain	16	1	1d2v, D113-D149
		variant surface glycoprotein MITat1.2, C-terminal domain	1	1	1xu6, A354-A433
27	β-hairpin and 1 α-helix disulfide-bonded to N-terminal loop	Kazal family serine protease inhibitor-like	76	21	1tbq, R1-R51
		plant serine protease inhibitor-like	16	9	1ce3, A1-A54
28	folded hairpin	ATI-like serine protease inhibitor	9	5	1ccv, A1-A56
29	folded and twisted hairpin	BPTI-like serine protease inhibitor/ dendrotoxin-like	156	28	1aap, A1-A56
30	4-strand β-sheet, antiparallel, strand order 1234 and 1 α-helix	neurophysin II	22	5	1npo, A5-A52
31	4-strand β-sheet, antiparallel, strand order 2134 and 2 α-helices	TGFβ binding protein-like	7	4	1uzj, B2530-B2606
32	4-strand β-barrel, strand order 1243, and 1 α-helix	hydrophobin II	2	1	1r2m, A1-A70
33	1 α-helix and 4 β-strands in flat array, meander topology	thyroglobulin-like domain	6	3	1l3h, A1-A65
34	2 β-hairpins and 2 α-helices	TIMP, C-terminal subdomain	7	3	1gxd, C120-C192
35	small disulfide-closed loop	μ/α conotoxin-like	46	28	1tcg, 1-22
		mini-protein 2 (synthesized)	13	5	1hqq, E1-E13
		guanylin/ heat-stable enterotoxin-like	7	5	1uya, 1-16
		orexin A	2	1	1wso, A1-A33
		arylsulfatase, conotoxin-like subdomain	8	3	1auk, 487-503
36	mostly coil, N-terminal α-helix	minicollagen-I, C-terminal domain	2	1	1sop, A1-A24
37	mostly coil, central α-helix	penaeidins	2	2	1ueo, A1-A63
38	mostly coil, C-terminal α-helix	tertiapin	1	1	1ter, 1-21
39	mostly coil, 1 small α-helix	somatomedin B domain	3	2	1s4g, A1-A51
40	mostly coil, left-handed loop followed by right-handed loop	LDL receptor-like domain	20	17	1ajj, 4-40
41	mostly coil, right-handed	sea anemone neurotoxin III	1	1	1ans, 1-27

In addition to this general structural similarity, two of these families, the CRISP (cysteine-rich secretory protein) family helical bundle subdomain and sea anemone toxin K, share similar disulfide-bonding patterns and have an N-terminal extension that is disulfide-bonded to the third α-helix in the bundle (Figure 3.1b). Representatives of these families (pdb|1bgk 1-37, *Bunodosoma granulifera* toxin (Dauplais, Lecoq et al. 1997);

pdb|1rc9 A181-A221, *Trimeresurus stejnegeri* stecrisp C-terminal domain (Guo, Teng et al. 2005)) superimpose with an RMSD of 2.0 Å over 31 C_α atoms. Additionally, it has been demonstrated that some members of the CRISP family inhibit a variety of different ion channels including voltage-gated calcium channels (Nobile, Noceti et al. 1996), calcium-activated potassium channels (Wang, Shen et al. 2005), cyclic nucleotide-gated ion channels (Brown, Haley et al. 1999), and ryanodine receptors (Morrissette, Kratzschmar et al. 1995). Although it has not yet been directly established whether the helical bundle subdomain is the region of this protein responsible for the channel-blocking activity, this is an attractive hypothesis because the sea anemone toxin K family members perform a similar function of blocking potassium channels. Considering the high structural similarity and potential functional similarity, a homology relationship between these two families seems plausible. However, due to the low sequence similarity between these proteins (average identity ~13%) and the unconfirmed function of the CRISP family helical bundle subdomain, the merging of these two families cannot yet be confidently asserted.

Distribution of families within fold groups

Each of the 41 fold groups in this classification contains between 1 and 8 distinct families (Figure 3.1a). There is a subset of topologies that seem to be quite common among small, disulfide-rich domains. In fact, nearly half of the 98 families belong to fold groups that consist of 5 or more non-homologous families each. Examples of recurring structural motifs in disulfide-rich domains are depicted in Figure 3.1. Typically, these common folds have a simple topology that could easily have arisen multiple times by chance, such as α-hairpins (fold groups 1 and 2) or 3-strand β-sheets with meander topology (fold group 12). Knottin-like topology, found in nearly 40% of the disulfide-rich domain structures currently available (fold groups 23-25), is the most commonly observed structural motif. It is characterized by two adjacent disulfide bonds (one bond is formed by the 1st and 3rd cysteines in the primary sequence, and the other by the 2nd and 4th cysteine residues) and a conserved β-hairpin, on which the 3rd and 4th cysteines

Figure 3.1: Common Folds Adopted by Unrelated Disulfide-Rich Families

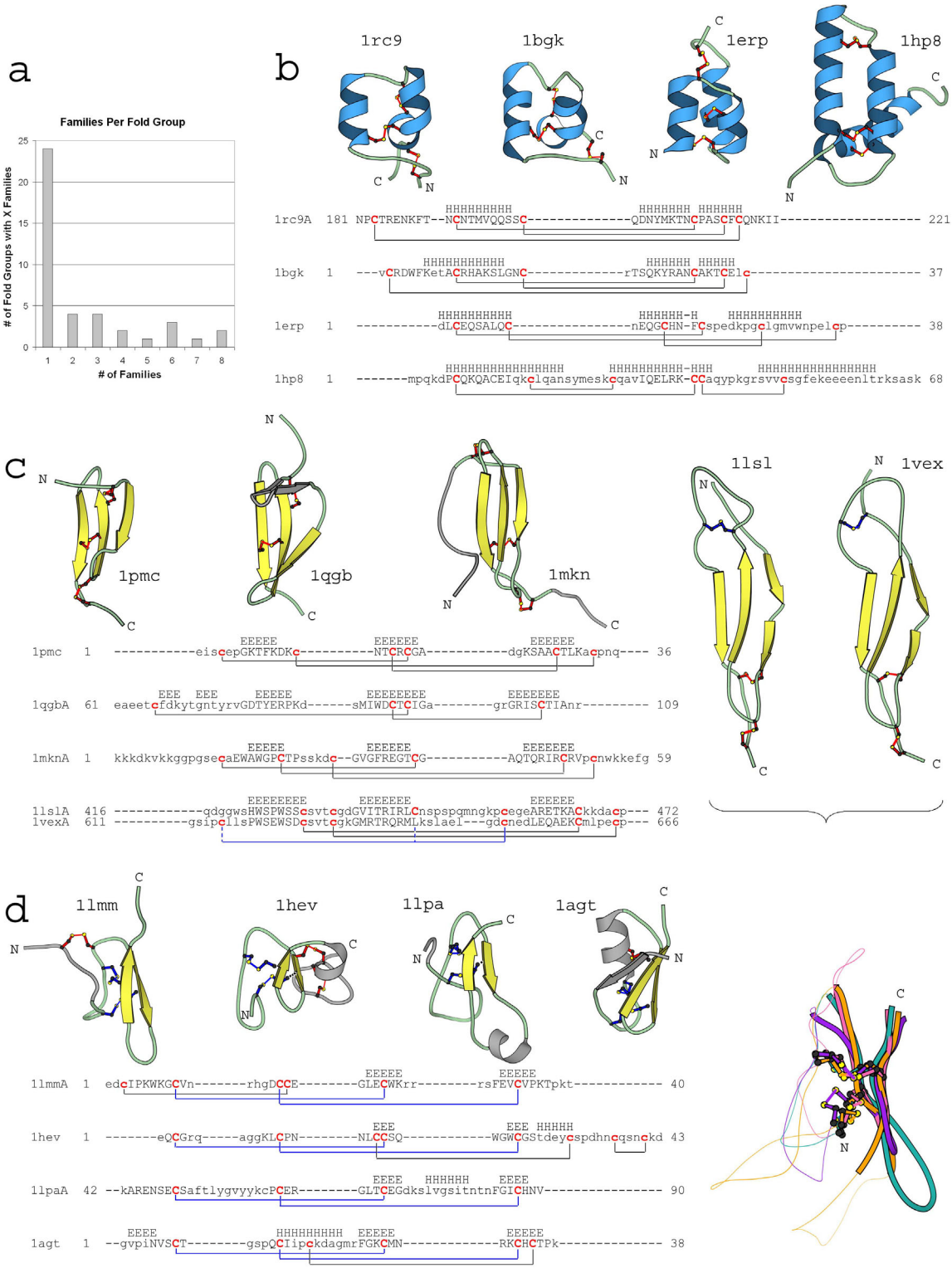


Figure 3.1: Common Folds Adopted by Unrelated Disulfide-Rich Families. The common structural core of each fold group is shown in color (blue α-helices, yellow β-strands, green coils); additional elements are shown in grey. In this and other MOLSCRIPT figures in Chapter 3, disulfide bonds are colored red or blue and shown in ball-and-stick format. A structural alignment for the representatives is shown below the structure figures. In these and other alignments in Chapter 3, capital letters denote residues that are structurally aligned; lower case letters are residues that do not align structurally with the other fold group members. Red bold text indicates cysteine residues that are highly conserved within a family and are involved in disulfide bonds. PDB identifiers and chain names are shown at the far left. The numbers before and after the sequence denote the PDB residue number of the first and last residues in the domain sequence, respectively. Secondary structure elements are noted above the alignment with H signifying α-helix and E signifying β-strand. Disulfide bonding patterns are depicted by lines connecting cysteine residues. a) Distribution of disulfide-rich families in fold groups. b) Representatives of four families with right-handed 3-helix bundle fold (fold group 3): *Trimeresurus stejnegeri* stecrisp helical subdomain (pdb|1rc9), *Bunodosoma granulifera* toxin K (pdb|1bgk), *Euplotes raikovi* pheromone ER-10 (pdb|1erp), and *Homo sapiens* p8 protein from oncogene MTCP1 (pdb|1hp8). c) Representatives of four families with antiparallel 3-stranded β-sheet with meander topology (fold group 12): *Locusta migratoria* protease inhibitor PMP-C (pdb|1pmc), *Homo sapiens* fibronectin type 1 module (pdb|1qgb), *Homo sapiens* midkine N-terminal domain (pdb|1mkn), and TSP-1 repeats from *Homo sapiens* thrombospondin (pdb|1lsl) and *Rattus norvegicus* F-spondin (pdb|1vex). The bracket indicates homologous domains. Disulfide bonds shown in blue involve a cysteine that does not align among all members of that family. d) Representatives of four families with knottin-like topology (fold group 23): *Psalmopoeus cambridgei* psalmotoxin 1 (pdb|1lmm), *Hevea brasiliensis* hevein (pdb|1hev), *Sus scrofa* colipase C-terminal domain (pdb|1lpa), and *Leiurus quinquestriatus hebraeus* agitoxin (pdb|1agt). Disulfide bonds shown in blue form the distinctive disulfide cross. The key knottin-like features (disulfide cross and β-hairpin) are superimposed for these four representatives (pdb|1lmm, green; pdb|1hev, purple; pdb|1lpa, orange; pdb|1agt, pink). The connecting backbone is shown as a thin coil.

---

are located (Figure 3.1d). Because the two disulfide bonds are roughly perpendicular so that they form an "X" or cross, the knottin-like core is also known as the disulfide β-cross. This motif has previously been suggested as a stable protein folding nucleus (Harrison and Sternberg 1996), which would confer an evolutionary advantage to proteins with this particular fold and explain the convergence of a large number of families to this common core.

On the other hand, approximately half of the 41 fold groups currently include only a single protein family. Some of these proteins have more complicated architectures (for example, irregular α-helical arrays; fold groups 5-8), while others are mostly-coil proteins with little or no standard secondary structure (fold groups 36-41). This large number of unique folds reflects the wide conformational variety available to proteins stabilized by disulfide bonds. Because a structural scaffold that is not entirely reliant upon secondary structure and hydrophobic interactions allows for much more conformational irregularity

(e.g. little or no α-helix and β-strand character, non-globular shapes, etc.), disulfide bonds can potentially stabilize numerous protein conformations that otherwise would not exist. For example, *Ascaris suum* chymotrypsin/elastase inhibitor (fold group 28) is a 60-residue protein with a structural fold maintained by 5 conserved disulfide bonds. As this protein contains only a few small secondary structure elements and lacks a hydrophobic core (Huang, Strynadka et al. 1994), it is highly unlikely that a non-disulfide structural analog of this fold would exist in nature.

It should also be noted that the currently available disulfide-rich domain structures are not expected to represent all disulfide-stabilized proteins that exist in nature. It is likely that this classification will require the addition of several new families and fold groups when novel disulfide-rich protein structures are revealed in the future.

3.3.2 Comparison to SCOP database

Most of these disulfide-rich domains are also classified by the SCOP database. Approximately 13% of the 2945 domains are not assigned a SCOP classification (version 1.69), in most cases because the structure of the protein was released quite recently and has not yet been incorporated into the SCOP database. Another 4% of domains in this study are regions within larger proteins that are not distinguished as distinct subdomains by SCOP (for example, the N-terminal EGF-like domain of alliinase; pdb|1lk9 (Kuettner, Hilgenfeld et al. 2002a), residues A2-A60). Of the remaining 2446 domains (i.e. those that are classified by SCOP), 84% are found in the "Small proteins" class of SCOP, 11.5% in the "all-α" class, 3% in the "Peptides" class, 1% in the all-β class, and the remaining 0.5% in the "Coiled coil proteins" and "Designed proteins" classes.

Fold group level describes broader structural similarity than SCOP folds

The SCOP fold level is designed to reflect strong structural similarity, where members of a common fold "have the same major secondary structures in the same arrangement and with the same topological connections" (http://scop.mrc-

lmb.cam.ac.uk/scop/intro.html). The fold group level in the classification presented here is comparable, in that domains within a fold group have a common structural core, but describes more broad similarities between protein structures. For example, *Eurplotes raikovi* pheromone ER-10 (pdb|1erp (Brown, Mronga et al. 1993)) and mature T-cell proliferation oncogene-encoded protein $p8^{MTCP1}$ (pdb|1hp8 (Barthe, Yang et al. 1997)) are two unrelated small α-helical proteins. Because the structures of both proteins are right-handed 3-helix bundles, they are assigned the same fold group in this work (fold group 3; Figure 3.1b). However, SCOP assigns these proteins to separate folds (a.10: "protozoan pheromone proteins" and a.17: "p8-MTCP1"), most likely because of the different sizes and relative orientations of the α-helices. The broad nature of the fold groups is reflected by the distribution of SCOP folds in this classification: 18 of the fold groups presented here include proteins from more than one SCOP fold, and 6 of the fold groups include representatives of more than one SCOP class.

The one exception to this trend is the set of knottin-like domains. SCOP assigns all of these domains to the same fold, g.3: "Knottins (small inhibitors, toxins, lectins)", based on the presence of the two adjacent disulfide bonds and β-hairpin. By fold group definition used in this study, however, all secondary structure elements in the structural core of a domain should be considered. Therefore, in the present classification, knottin-like structures are arranged according to the topology of the backbone contributing all four cysteine residues that make up the disulfide cross, rather than only the β-hairpin that contains the 3rd and 4th cysteines. The knottin-like domains comprise fold groups 23, 24, and 25 (Table 3.2).

Disulfide-rich families are approximately equivalent to SCOP superfamilies

The disulfide-rich families presented in this work are, for the most part, consistent with the superfamily level of SCOP, which is the broadest level of homology conveyed in the SCOP classification hierarchy. SCOP, however, is a fairly conservative database and after careful examination of the domains in this current study, a few additional homology relationships were identified. These newly linked families are described in section 3.3.3.

3.3.3 Distant homology between disulfide-rich domains

Bowman-Birk inhibitors and bromelain inhibitors

Bowman-Birk inhibitors (BBIs) are serine protease inhibitors specific for trypsin and chymotrypsin. These proteins are found in many plant seeds, and structures have so far been solved for BBIs from soybean, lima bean, mung bean, adzuki bean, garden pea,

Figure 3.2: Bowman-Birk and Bromelain Inhibitors

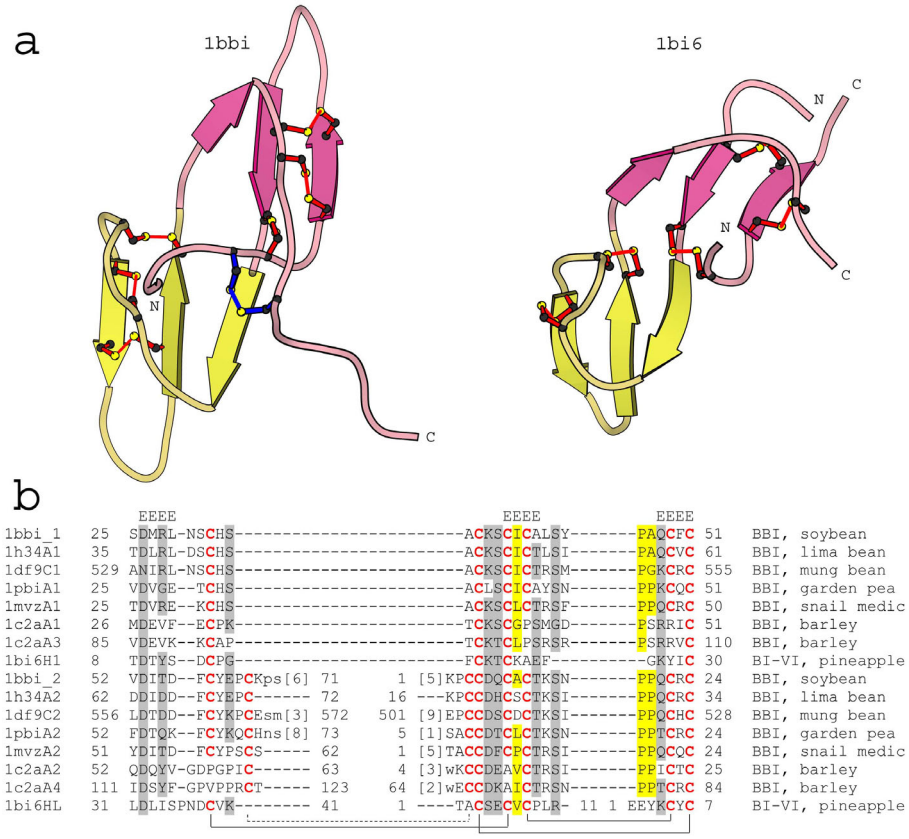


Figure 3.2: Bowman-Birk and Bromelain Inhibitors. a) MOLSCRIPT diagrams of Bowman-Birk inhibitor from soybean (pdb|1bbi) and bromelain inhibitor VI from pineapple (pdb|1bi6). Each protein is comprised of two homologous domains that adopt an antiparallel 3-stranded β-sheet fold; one domain is continuous (yellow) while the other is circularly permuted (pink). Highly conserved disulfide bonds are shown in red; additional disulfide bonds are shown in blue. b) Structure-based multiple alignment of BBI and BI-VI representatives. Solid lines indicate highly conserved disulfide bonds (red in panel a); dashed lines indicate disulfide bonds that are found in only one of the two tandem repeats (blue in panel a). In this and other multiple alignments in Chapter 3, red bold text indicates conserved cysteine residues involved in disulfide bonds, yellow highlighting indicates uncharged residues in mostly hydrophobic positions, grey highlighting indicates mostly polar positions, and cyan highlighting indicates mostly aromatic positions.

barley, peanut, and snail medic seeds. Also, seven isoinhibitors of the cysteine protease bromelain have been identified from the stem of pineapple. The structure of bromelain inhibitor VI (BI-VI) is currently available. The bromelain inhibitors are somewhat unique in that they are formed by a heavy chain (41 residues) and a light chain (11 residues) which originated from a single-chain precursor (Hatano, Kojima et al. 1995; Sawano, Muramatsu et al. 2002). The BBI and BI-VI proteins have very clear structural similarity (Figure 3.2a). First, both proteins contain a tandem repeat of a small, antiparallel 3-stranded β-sheet with strand order 231 and 3 highly conserved disulfide bonds. In both BBIs and BI-VI, one domain is contiguous while the second domain is related by circular permutation. Furthermore, while the sequence similarity between these proteins is not overwhelming, the percent identity between the BBI and BI-VI domains is comparable with the identity between the two subdomains of the same protein. For example, the average identity between the circularly permuted domain of BI-VI (1bi6HL in Figure 3.2) and the BBI representative domains is 27%, while the identity between 1bi6HL and the non-permuted BI-VI domain (1bi6H1) is only marginally higher at 28%. Furthermore, BBIs and BI-VI are identified as homologs by the MEROPS database of proteases and protease inhibitors (clan IF) (Rawlings, Tolle et al. 2004). Based on the sequence, structural, and functional similarity between these proteins, the Bowman-Birk inhibitors and bromelain inhibitor VI are merged into a single family, which is included in fold group 13.

EGF-like subdomain of garlic alliinase

Garlic alliinase is a lyase that cleaves carbon-sulfur bonds to produce the sulfur-containing garlic components to which this plant's pharmacological properties are attributed. The structure of this protein confirmed the predicted presence of a cysteine-rich N-terminal subdomain similar to EGF-like domains (Kuettner, Hilgenfeld et al. 2002a; Kuettner, Hilgenfeld et al. 2002b). The function of this subdomain is currently unknown. This domain lacks one of the three highly conserved disulfide bonds that are typical of EGF-like domains, and has one additional disulfide bond that is not seen

Figure 3.3: EGF-like Subdomain of Garlic Alliinase

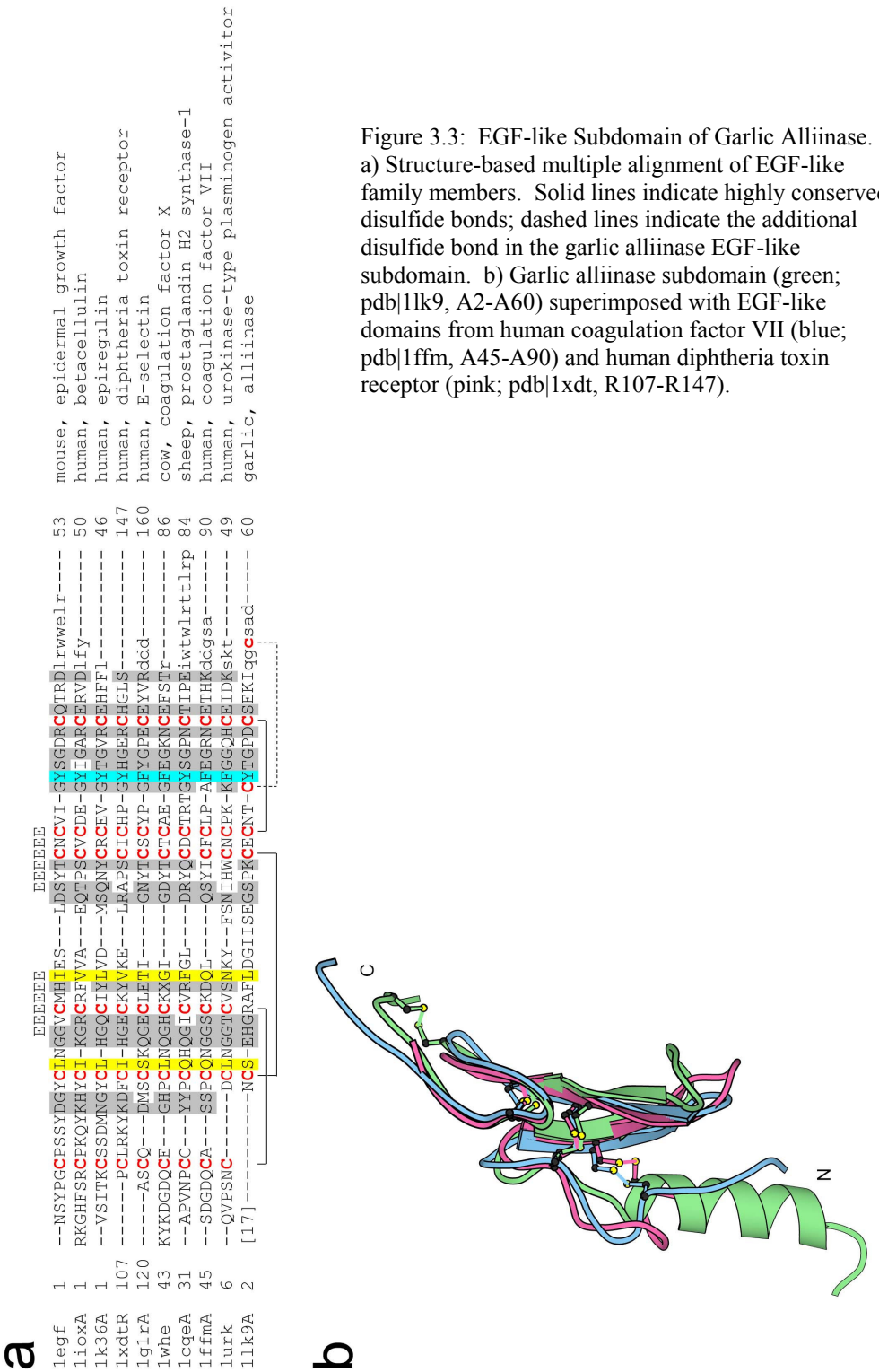


Figure 3.3: EGF-like Subdomain of Garlic Alliinase. a) Structure-based multiple alignment of EGF-like family members. Solid lines indicate highly conserved disulfide bonds; dashed lines indicate the additional disulfide bond in the garlic alliinase EGF-like subdomain. b) Garlic alliinase subdomain (green; pdb|1lk9, A2-A60) superimposed with EGF-like domains from human coagulation factor VII (blue; pdb|1ffm, A45-A90) and human diphtheria toxin receptor (pink; pdb|1xdt, R107-R147).

among other family members (Figure 3.3a). Despite these differences in disulfide bonding patterns, the alliinase N-terminal domain nonetheless shares striking structural similarity with the EGF-like family. Among the closest structural neighbors of the alliinase subdomain are EGF-like domains from human coagulation factor VII (pdb|1ffm; 2.1 Å RMSD, 33 C_α atoms) and human diphtheria toxin receptor (pdb|1xdt; 2.1 Å RMSD, 31 C_α atoms) (Figure 3.3b). Although there is limited sequence similarity between the alliinase subdomain and the other EGF-like family members (typically <20% identity), this distant homology relationship is also recognized by the Pfam database (Bateman, Coin et al. 2004), which places the N-terminal domain of garlic alliinase (PF04863) into the same clan as other EGF-like domains (CL0001: EGF superfamily). The N-terminal subdomain of garlic alliinase has therefore been included with the family of EGF-like knottins in fold group 25.

Cellulose binding/docking domains

The third example includes domains from two different polysaccharide hydrolases that are involved in recycling carbohydrates by breaking down plant cell walls. The cellulose binding domain of *Pseudomonas fluorescens* xylanase A (CBDx) binds carbohydrate polymers of cell walls (Millward-Sadler, Davidson et al. 1995). The cellulose docking domain of *Piromyces equi* endoglucanase Cel45A (CDDe) does not interact directly with cellulose but instead binds to small protein domains (cohesin domains) which are found in the same polypeptide chain as domains that do directly bind cellulose (Fanutti, Ponyi et al. 1995). Thus, while CBDx and CDDe perform different molecular functions, they are responsible for the same role, on a more general level, of bringing the catalytic domains of these enzymes and the carbohydrate polymers into close proximity so that the hydrolysis reaction can proceed. The structural features shared by these domains include two hairpins that lie approximately perpendicular to each other (Figure 3.4a). This region of these domains superimposes with 2.7 Å RMSD (25 C_α atoms). Additionally, the key residues involved in binding are presented on the same face of both structures (Ponyi, Szabo et al. 2000; Raghothama, Eberhardt et al. 2001)

(Figure 3.4a). Although these domains share only limited sequence similarity, the distant homology relationship between them is also identified by the Pfam database (PF02013: cellulose or protein binding domain). Thus, the CBDx and CDDe are assigned to the cellulose binding/docking domain family (fold group 18).

Figure 3.4: Cellulose Binding/Docking Domains

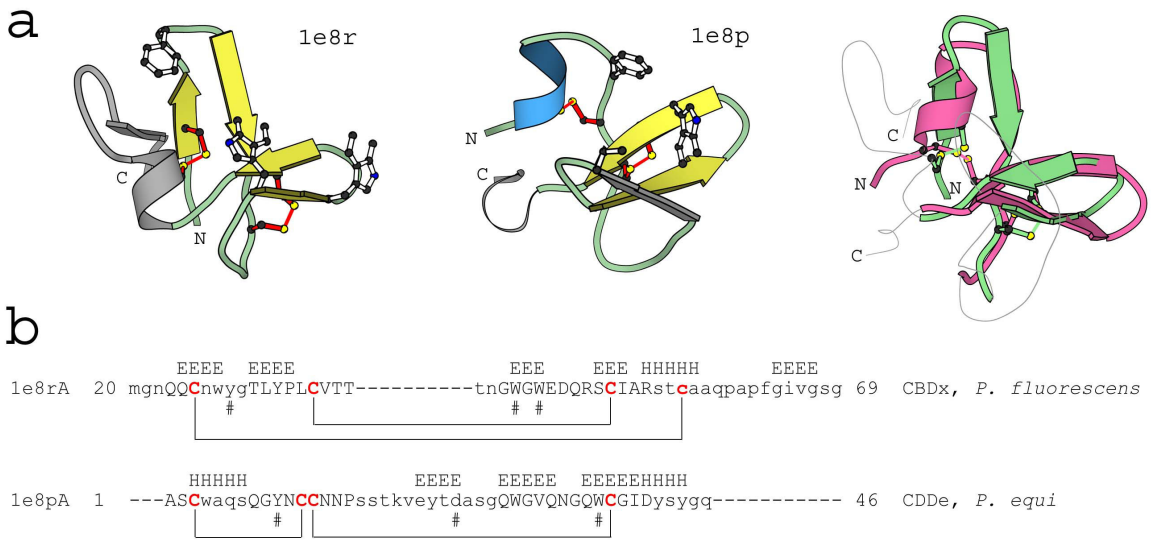


Figure 3.4: Cellulose Binding/Docking Domains. a) The cellulose binding domain of xylanase A (CBDx) from *Pseudomonas fluorescens* (pdb|1e8r) and the cellulose docking domain of endoglucanase Cel45 (CDDe) from *Piromyces equi* (pdb|1e8p) share a similar structural core of two roughly perpendicular hairpins. Shared structural features are shown in color: CBDx has two β-hairpins, while CDDe has one β-hairpin and one hairpin formed by an α-helix and a coil region. Other elements are shown in grey. Putative binding residues are shown in ball-and-stick format and colored white. In the superimposed view of CBDx (green) and CDDe (pink), the two hairpins are shown in color and the rest of the backbone is shown as thin grey coil. b) Structural alignment of the CDDe and CBDx domains. Putative binding residues are indicated by #.

Knottin-like domains

The knottin-like proteins are a large group of structurally similar proteins, as described in section 3.3.1. SCOP assigns these domains to 19 superfamilies in a single fold. The classification presented here also includes 19 knottin-like families, although some of these proteins are found outside of the SCOP knottin-like fold. Furthermore,

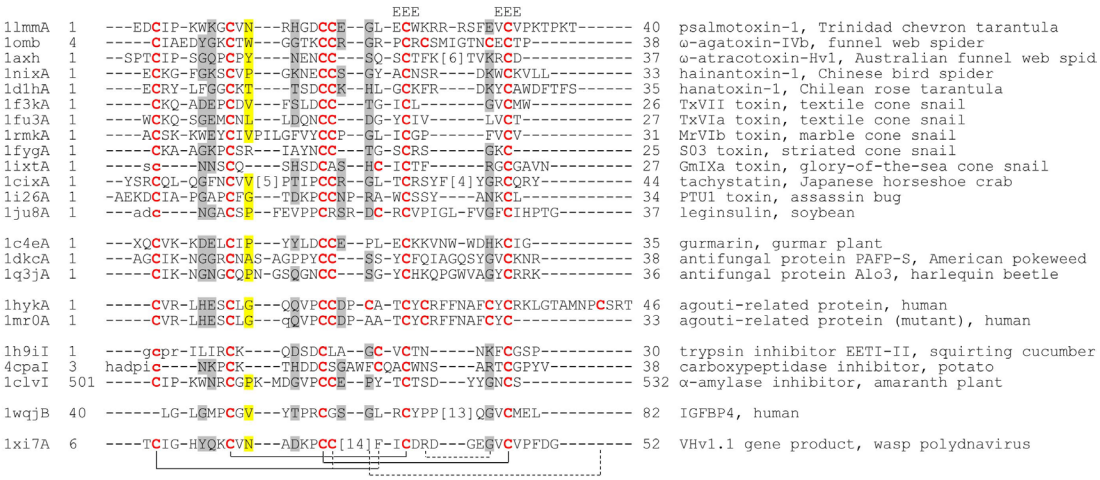
while most of the knottin-like families in this current study are in close agreement with SCOP, subtle rearrangements of some families within this large class have been made.

The omega toxin-like superfamily is among the largest knottin-like superfamilies in SCOP. While most of the members are ω-conotoxins and spider toxins, this superfamily also includes some insect toxins, scorpion toxins, spider lectins, and antimicrobial proteins. Henceforth, the omega toxin-like superfamily of SCOP will be referred to as the ωTL superfamily. This set of proteins makes up the bulk of one family in this classification (henceforth referred to as the spider toxin-like family), although proteins from several other SCOP superfamilies have also been included. These additional members include gurmarin, antifungal peptides PAFP-S and Alo3, the C-terminal domain of agouti-related signaling proteins, the knottin-like domain of insulin-like growth factor binding proteins (IGFBPs), plant enzyme (α-amylase, carboxypeptidase, trypsin) inhibitors, and the C-terminal subdomain of the VHv1.1 polydnaviral gene product. With the exception of the viral subdomain, which is not yet incorporated into the SCOP database, each of these domains is found outside of the ωTL SCOP superfamily, perhaps on the basis of functional dissimilarity. However, all of these new members display significant sequence and structural similarity with the ωTL domains (Figure 3.5b). Each of these additional subsets includes at least one domain that shares >33% sequence identity and <2.5 Å RMSD ($\geq$25 C_α atoms) with a member of the ωTL SCOP superfamily.

Furthermore, there are several cases in which these new members share greater sequence and structural similarity with a representative of the ωTL SCOP superfamily than with other members of its SCOP-assigned superfamily. For example, antifungal protein PAFP-S (pdb|1dkc) is highly similar to ωTL representative conotoxin TxVIa (pdb|1fu3) with 37% sequence identity and 2.1 Å RMSD (27 C_α atoms). Meanwhile, sweet-taste suppressor signaling protein gurmarin (pdb|1c4e) shares 38% sequence identity and 1.9 Å RMSD (26 C_α atoms) with ωTL representative conotoxin TxVII (pdb|1f3k). Although PAFP-S and gurmarin are assigned to the same SCOP superfamily, they share less similarity with each other (20% sequence identity; 3.2 Å RMSD, 33 C_α

Figure 3.5: Additional Members of the Spider Toxin-like Family

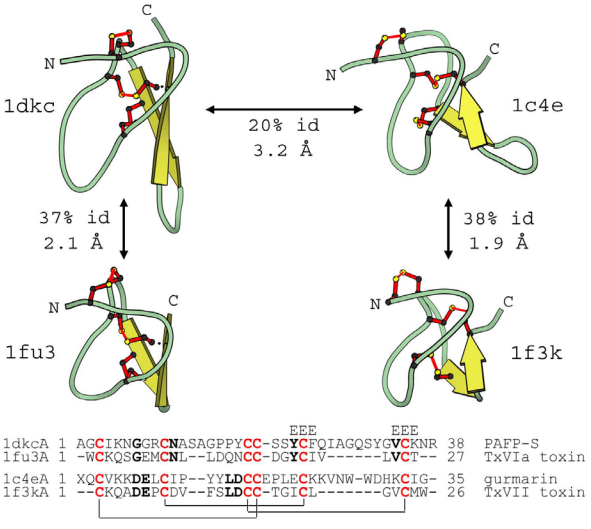
a



b

additional ωTL family members (SCOP superfamily)	*similarity to ωTL (g.3.6) domains*				
	sequence identity		*RMSD*		*linking pair example* (sequence identity, RMSD)
	average	*best*	*average*	*best*	
gurmarin and antifungal proteins (g.3.4)	26%	40%	2.7 Å	1.2 Å	gurmarin signaling protein (pdb|1c4e) & conotoxin TxVII (pdb|1f3k) 38%, 1.9 Å (26 Cα)
plant enzyme inhibitors (g.3.2)	24%	41%	2.3 Å	1.0 Å	α-amylase inhibitor (pdb|1clv) & covalitoxin-I (pdb|1v5a) 36%, 1.9 Å (27 Cα)
agouti-related, C-terminal domain (g.3.5)	24%	38%	2.7 Å	1.7 Å	agouti-related protein (pdb|1mr0) & hainantoxin-I (pdb|1nix) 33%, 2.1 Å (30 Cα)
IGFBP knottin-like domain (g.3.9)	21%	35%	2.3 Å	1.3 Å	IGFBP4 subdomain (pdb|1wqj, B40-B82) & conotoxin TxVII (pdb|1f3k) 35%, 2.5 Å (25 Cα)
C-terminal domain, VHv1.1 viral gene product (N/A)	25%	37%	2.8 Å	1.6 Å	viral subdomain (pdb|1xi7) & conotoxin TVIIa (pdb|1eyo) 37%, 1.6 Å (26 Cα)
similarity among ωTL domains	28%	--	2.4 Å	--	--

c



d

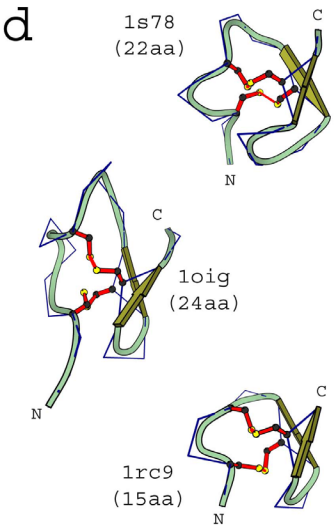


Figure 3.5: Additional Members of the Spider Toxin-Like Family. a) Structure-based multiple alignment of ωTL domains and added family members. Solid lines indicate highly conserved disulfide bonds; dashed lines indicate disulfide bonds found in few family members. b) Calculated similarity between ωTL domains and added family members. Average values were calculated using representatives at 95% identity. A "linking pair" is defined as a pair of domains from different SCOP superfamilies that share >33% sequence identity and <2.5 Å RMSD ($\geq$25 C_α). c) American pokeweed antifungal protein PAFP-S (pdb|1dkc) and gurmar plant signaling protein gurmarin (pdb|1c4e) are assigned to the same SCOP superfamily but share less similarity with each other than with ωTL conotoxins TxVIa (pdb|1fu3) and TxVII (pdb|1f3k), respectively. Boldface letters in the alignment indicate identical residues between the domains assigned to different SCOP superfamilies. d) Very small knottin-like domains: cysteine-rich repeat from human ErbB2 (pdb|1s78 A171-A192), DPY module from *Drosophila melanogaster* dumpy protein (pdb|1oig), and knottin-like subdomain of *Trimeresurus stejnegeri* stecrisp (pdb|1rc9 A166-A180). C_α traces are shown in order to clarify which β-strand contributes each "hanging" cysteine side chain.

---

atoms) than with the ωTL representatives (Figure 3.5c). Likewise, plant inhibitors of trypsin, carboxypeptidase A, and α-amylase inhibitor belong to the same SCOP superfamily ("plant inhibitors of proteinases and amylases") despite the limited sequence and structural similarity among these proteins. The α-amylase inhibitor shares only 16% sequence identity and 3.2 Å RMSD (24 C_α atoms) with carboxypeptidase A inhibitor, and only 22% sequence identity and 4.0 Å RMSD (27 C_α atoms) with the trypsin inhibitors. However, the α-amylase inhibitor (pdb|1clv) shares 36% sequence identity and 1.9 Å RMSD (27 C_α atoms) with ωTL representative covalitoxin-I (pdb|1v5a). In a similar example, *Conus gloriamaris* conotoxin GmIXa (pdb|1ixt), the first structural representative of the P-superfamily conotoxins (Miles, Dy et al. 2002), is assigned to the ωTL of SCOP, presumably on the basis of putative function and species of origin. However, this protein shares more obvious similarity with carboxypeptidase A inhibitor (pdb|4cpa) from the "plant inhibitors of proteinases and amylases" SCOP superfamily (41% sequence identity; 2.3 Å RMSD, 25 C_α atoms) than with the other ωTL domains (average sequence identity = 23%; average RMSD = 2.5 Å, 24 C_α atoms). Thus, there are members of ωTL SCOP superfamily with higher similarity to other SCOP superfamilies than to other ωTL domains, and there are members of other SCOP superfamilies with higher similarity to ωTL domains than to each other. On the basis of such links, 4 different SCOP superfamilies have been merged with the ωTL domains.

In most of these cases, entire SCOP superfamilies are merged with the ωTL domains. The sole exception is the knottin-like domain of IGFBP. In SCOP, this domain

is classified with the cysteine-rich repeats of EGF receptors (also known as ErbBs). The extracellular domain of ErbB proteins includes 2 regions of cysteine-rich repeats, each of which is comprised of 13 homologous knottin-like domains in tandem, and 2 regions of leucine-rich repeats. The cysteine-rich regions are involved in dimerization (Cho and Leahy 2002). On the other hand, 3 non-similar disulfide-rich domains are found within IGFBPs. The second domain adopts a knottin-like fold, and is involved in binding insulin-like growth factors (Kalus, Zweckstetter et al. 1998). The knottin-like domains of IGFBPs and ErbBs share only 19% sequence identity on average and are structurally alignable over only 12 C_α atoms, with an average RMSD of 2.1 Å over these residues. Thus, the knottin-like domains of IGFBPs and ErbBs share neither significant sequence, structural, nor functional similarity and are therefore unlikely to be evolutionarily related. The knottin-like domain of IGFBPs has been added to the spider toxin-like family. Meanwhile, the ErbB knottin-like repeats are very short (average size = 20 residues), and are most structurally similar to other very small knottin-like domains such as DPY modules and the knottin-like subdomain of CRISP family proteins (Figure 3.5d). However, due to the limited sequence similarity between these domains (average sequence identity of 25% between knottin-like domains of ErbBs and CRISPs; average sequence identity of 19% between ErbB knottin-like domains and DPY modules), they remain as 3 separate families within fold group 25.

3.4 DISULFIDE BONDING PATTERNS AND PROTEIN TOPOLOGY

3.4.1 Disulfide bonds and protein structure

While there is debate about the physical mechanism by which disulfide bonds act as a stabilizing influence (Flory 1956; Doig and Williams 1991; Betz 1993), the general importance of these covalent bonds in small protein domains is clearly demonstrated by their extremely high conservation. However, a comprehensive view of the numerous

cysteine-mutation studies performed on these proteins indicates that the extent to which a domain's structure and function are dependent upon the presence of conserved disulfide bonds varies. Mutagenesis studies of several different proteins have shown that eliminating a specific disulfide bond significantly alters neither the protein structure nor function: the mutated protein adopts a native-like fold (although is usually less stable and more susceptible to denaturation) and retains all or most of its wild type function (which is experimentally verified in some cases and hypothesized based on the lack of structural change in others). Small protein examples include bovine pancreatic trypsin inhibitor (Eigenbrot, Randal et al. 1990; van Mierlo, Darby et al. 1991; Perona, Tsu et al. 1993) (fold group 29), charybdotoxin (Song, Gilquin et al. 1997) (scorpion toxin-like/insect and plant defensin-like family, fold group 23), kalata B1 (Daly, Clark et al. 2003) (kalata-like cyclotides family, fold group 23), and vascular endothelial growth factor (Muller, Heiring et al. 2002) (cysteine knot cytokines family, fold group 23).

In other studies, however, elimination of a single disulfide bond resulted in drastic changes in protein structure and/or function. For example, mutating the disulfide bond between the first and third cysteines (i.e., the 1-3 bond) in murine epidermal growth factor (EGF-like family, fold group 25) results in a structural fold that is highly similar to the native fold except at the N-terminal tail, but causes a dramatic reduction in both mitogenic activity and receptor binding (Barnham, Torres et al. 1998). The opposite situation is seen upon mutation of the 2-3 disulfide bond in an endothelin-1 analog (endothelin-like family, fold group 1): the protein retains agonist activity but the native tertiary fold is completely destroyed (Hewage, Jiang et al. 1999). Eliminating the 2-4 disulfide bond of α-conotoxin GI (μ/α conotoxin-like family, fold group 35) results in both a non-native structural fold and loss of toxicity (Mok and Han 1999). A similar disruption of both structure and function is seen when the 2-4 or 3-5 disulfide bonds of toxin ShK (sea anemone toxin K family, fold group 3) are deleted. Interestingly, the ShK variant with a mutated 1-6 disulfide bond retains potassium channel inhibitor activity despite adopting a structure significantly different from the native fold(Pennington,

Lanigan et al. 1999). Unsurprisingly, very small proteins (<35 residues) tend to be highly intolerant to the mutation of cysteine residues involved in disulfide bond formation.

3.4.2 Native variations in disulfide bonds

It has been long since recognized that the cysteine residues in disulfide bonds are nearly always conserved as pairs (Thornton 1981). That is to say, loss of a disulfide bond in a protein is typically due to mutation of both contributing cysteine residues rather than only one. However, cases in which some members of a disulfide-rich family have fewer (or more) disulfide bonds relative to others are not uncommon. Potassium channel inhibitor conkunitzin-S1 (pdb|1yl2 (Bayrhuber, Vijayan et al. 2005)) is a member of the BPTI-like/dendrotoxin-like family (fold group 29). The structure is essentially identical

Figure 3.6: Disulfide-Rich Family Members with Lost Disulfide Bonds

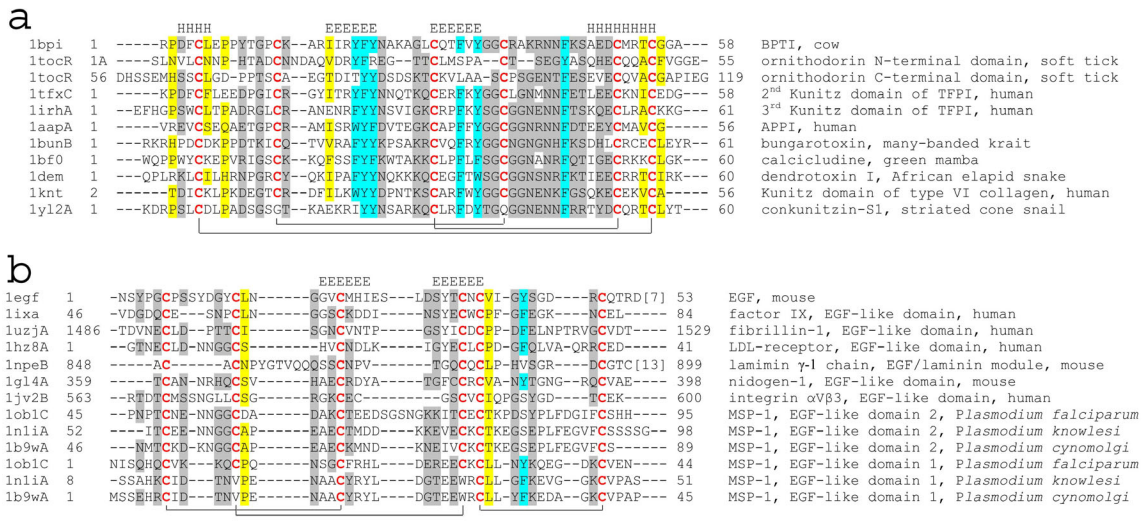


Figure 3.6: Disulfide-Rich Families Members with Lost Disulfide Bonds. a) BPTI-like/Kunitz/ Dendrotoxin-like family (fold group 29). Conkuntizin-S1 has only two of three highly conserved disulfide bonds. Protein abbreviations are as follows: bovine pancreatic trypsin inhibitor (BPTI), tissue factor pathway inhibitor (TFPI), and Alzheimer's amyloid B-protein precursor (APPI). b) EGF-like family (fold group 25). The N-terminal domain of MSP-1 from some *Plasmodium* species includes only two of the three highly conserved disulfide bonds. Protein abbreviations in this figure are as follows: epidermal growth factor (EGF), low density lipoprotein (LDL), and merozoite surface protein 1 (MSP-1). Details pertaining to alignment layout are described in Figure 3.2 legend.

to other family members (closest structural neighbor: Kunitz-type domain from human type VI collagen, pdb|1knt, 0.96 Å RMSD, 55 C_α atoms), despite that conkunitzin-S1 contains only two of the three disulfide bonds that are otherwise highly conserved in the BPTI-like family (Figure 3.6a). Similarly, the C-terminal region of merozoite surface protein 1 contains a tandem repeat of EGF-like domains, but the first repeat in some *Plasmodium* species has only two of the three highly conserved disulfide bonds of the EGF-like domain family while the second repeat has all three (Garman, Simcoke et al. 2003) (Figure 3.6b).

Numerous examples are also seen in which a few proteins have additional disulfide bonds relative to the majority of family members. Some spider toxins (for example ω-agatoxin IVa, pdb|1iva (Reily, Holub et al. 1994) or μ-agatoxin I, pdb|1eit (Omecinsky, Holub et al. 1996)) have an additional disulfide bond on the key β-hairpin of

Figure 3.7: Disulfide-Rich Family Members with Additional Disulfide Bonds

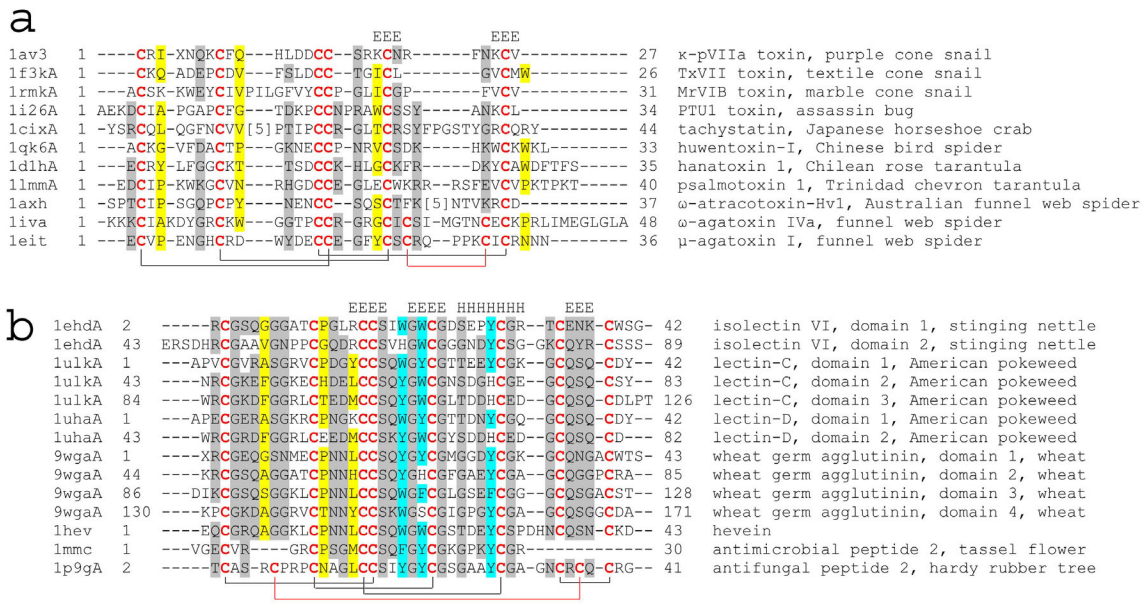


Figure 3.7: Disulfide-Rich Families Members with Additional Disulfide Bonds. a) Spider toxin/ω-conotoxin-like family (fold group 23). ω-agatoxin IVa and μ-agatoxin I have a disulfide bond not seen in all family members. b) Plant lectin-like family (fold group 23). Antifungal peptide 2 from hardy rubber tree has one additional disulfide bond that is currently unique to this particular protein structures. Details pertaining to alignment layout are described in Figure 3.2 legend.

their knottin-like fold (Figure 3.7a). Likewise, antifungal peptide 2 of hardy rubber tree (pdb|1p9g (Xiang, Huang et al. 2004)) contains a fifth disulfide bond that is not seen any other structures of plant lectin-like family members (fold group 23) (Figure 3.7b). Why certain family members seem to require fewer (or more) disulfide bonds than their homologs is not clear. Potential explanations might include different functional constraints, or different folding pathway requirements resulting from sequence variations between family members or the environment in which the organisms thrive.

Less common variations within families are seen when a cysteine in a disulfide bond is contributed from similar spatial positions but different regions of the protein sequence. One such example of a migrated cysteine is found in the thrombospondin type 1 family of fold group 12. Members of this family have three disulfide bonds, two of which are conserved in the sequence (i.e. formed by cysteine residues that align by sequence). The third disulfide bond (shown in blue in Figure 3.1c) is formed by the third and fourth cysteines in thrombospondin (TSP) (pdb|1lsl (Tan, Duquette et al. 2002)) but by the first and fourth cysteines in F-spondin (pdb|1szl, 1vex (Paakkonen, Tossavainen et al. *To be published*)). Although these residues (third cysteine of TSP and first cysteine of F-spondin) are separated by ~25 amino acids in the sequence, they are located in approximately equivalent spatial locations. When the TSP and F-spondin domains are superimposed (average RMSD: 3.4 Å, 49 C_α atoms), the S atoms of the migrated cysteine residues are ~4.2 Å apart. Cases such as these are intriguing because they suggest that maintaining the fold of a particular family (in this case, a very oblong 3-strand meander β-sheet) may require additional stabilization in a specific region of the structure.

More generally among fold groups, however, examples of shared disulfide-bonding requirements are not seen. Families within a fold group are often structurally stabilized by different numbers of disulfide bonds which cross-link different pairs of structure elements (for example, Figure 3.1b,c,d). Variations among bonding patterns suggests that while these domains do require disulfide bonds to maintain the protein fold, the specific arrangement of those bonds within the structure may not be particularly important. In fact, of the 17 fold groups in this classification that include more than one

family, the only cases in which all members share the same disulfide bonding patterns are the simple α-hairpins and β-hairpins.

3.4.3 Homologs with different disulfide bonding patterns

Among the most interesting examples are homologous or even identical proteins shown to have different disulfide bonding patterns. The family of disintegrins (fold group 21) includes proteins from *Viperidae* and *Crotalidae* snake venoms which inhibit biological processes such as platelet aggregation and tumor invasion by binding to integrins of the β1 and β3 classes (Gould, Polokoff et al. 1990). Despite their high sequence similarity (typically >40% identity among family members), the solved structures of representative disintegrins have revealed that these proteins have quite different topologies and disulfide bonding patterns (Figure 3.8a). These representatives can be divided into four groups based on disulfide connectivities. Grouping these proteins based on similarity of structural fold roughly parallels the groupings by disulfide bonding patterns. Members of the kistrin/trimestatin/flavoridin subset superimpose reasonably well with schistatin and three novel *Echis carinatus* disintegrin polypeptides (average RMSD between the subsets: 2.31 Å, 61 C_α atoms), which is not unexpected considering the four disulfide bonds they have in common. Conversely, echistatin and obtustatin also have four disulfide bonds in common, as well as nearly 50% sequence identity, but have quite different folds. Obtustatin has a compact, globular shape while echistatin adopts a more extended conformation similar to the other disintegrin structures. Salmosin is unlike all other disintegrins with solved structure both in structural fold and disulfide bonding pattern. The most conserved structural feature among disintegrins is a β-hairpin containing the RGD motif, which is involved in binding integrin. Although the disintegrins all perform the same general function (integrin-binding), they are relatively selective for different integrin-ligand interactions (Calvete, Marcinkiewicz et al. 2005). Additionally, some of these proteins function as homodimers (schistatin (Bilgrami, Tomar et al. 2004)), others as heterodimers (*Echis carinatus* novel disintegrin (Bilgrami,

Yadav et al. 2005)), and still others as monomers (kistrin (Adler, Lazarus et al. 1991)). It has been suggested that the different integrin inhibition specificities may result from the variations in surface change distribution or the striking conformational differences observed between family members (Paz Moreno-Murciano, Monleon et al. 2003; Shin, Hong et al. 2003). This family is intriguing in that despite such clear sequence and functional similarity, the disintegrins exhibit significant variations in disulfide bonding patterns, structural fold, and dimerization state.

In a related example, different disulfide bonding patterns are seen for the same protein domain. The somatomedin B (SMB) domain of human vitronectin, an adhesive glycoprotein found in blood, contains binding sites for plasminogen activator inhibitor type-1 (PAI-1), urokinase-type plasminogen activator receptor, and integrins. The three structures currently available of the SMB domain all have different disulfide bonding patterns (Zhou, Huntington et al. 2003; Kamikubo, De Guzman et al. 2004; Mayasundari, Whittemore et al. 2004) (Figure 3.8b). Interestingly, it was observed that several alternative bonding patterns would be compatible with the same fold (Kamikubo, De Guzman et al. 2004). For example, two of the SMB domain structures in the PDB (pdb|1oc0 and pdb|1ssu) superimpose with 2.0 Å RMSD over 36 C_α atoms despite having only one of four disulfide bonds in common. Furthermore, it was demonstrated that the PAI-1 binding function of this domain was retained by these dissimilar folds. As the only shared feature of these folds was a short α-helix containing the previously identified key functional residues (Figure 3.8b), it was suggested that function is maintained because each of the disulfide bonding patterns is compatible with the formation of this essential secondary structure element (Mayasundari, Whittemore et al. 2004). While only one native bonding pattern apparently exists in human blood (*aabcdbcd*) (Horn, Hurst et al. 2004; Mayasundari, Whittemore et al. 2004), it is nonetheless interesting to note the dramatic variations in global fold and disulfide bonding patterns that are tolerated by this domain without sacrificing function.

Figure 3.8: Variations in Disulfide-Bonding Patterns Within Families

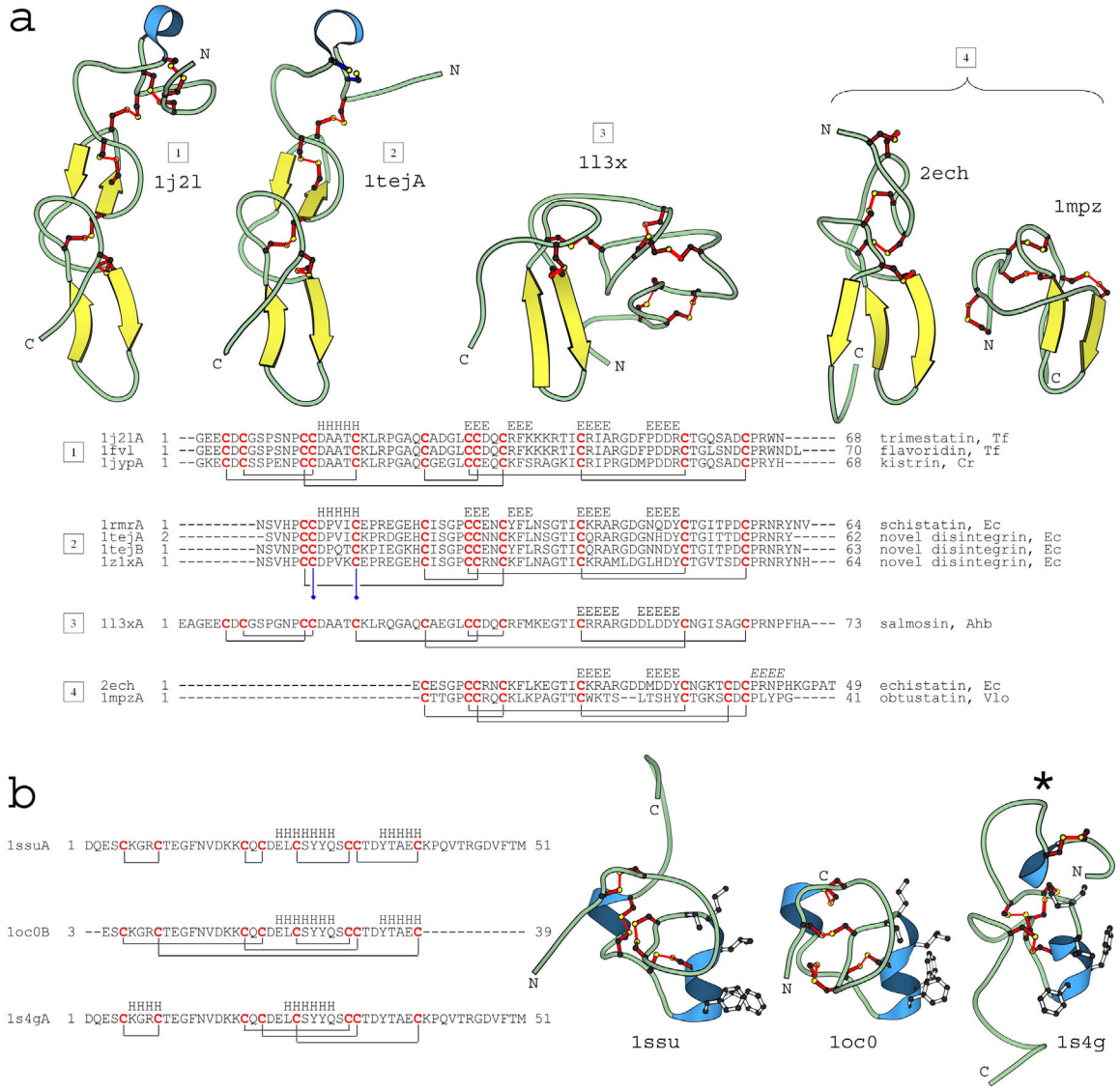


Figure 3.8: Variations in Disulfide-Bonding Patterns Within Families. a) Disintegrins with solved structures are grouped into four sets based on disulfide bonding patterns. In some (but not all) cases, similar bonding patterns result in similar topologies. Cysteines involved in intramolecular disulfide bonds are shown in red; cysteines involved in intermolecular disulfide bonds are shown in blue. In the alignment, secondary structure elements not conserved in all proteins of a subset are shown in italics. Intramolecular disulfide bonds are shown as black lines; intermolecular disulfide bonds are shown as blue diamond arrows. Species abbreviations are as follows: Tf *Trimeresurus flavoviridis*, Cr *Calloslasma rhodostoma*, Ec *Echis carinatus*, Ahb *Agkistrodon halys brevicaudus*, Vlo *Vipera lebetina obtusa*. b) Three structures of the SMB domain of human vitronectin have different disulfide bonding patterns. Functional residues are shown in ball-and-stick format. The native bonding pattern is indicated by an asterisk.

3.4.4 Disulfide bonding patterns observed in small protein domains

Occurrences of disulfide bonding patterns in proteins were previously analyzed by Benham and Jafri (Benham and Jafri 1993). In their study, the bonding patterns observed in 186 non-identical protein chains from the PDB structure database and the National Biomedical Research Foundation protein sequence database (now a part of UniProt (Apweiler, Bairoch et al. 2004)) were specified and evaluated in terms of two intrinsic properties: symmetry and reducibility. This analysis has been repeated with the domains in the present classification, as described in section 3.2.3 (Table 3.4).

Table 3.4: Bonding Patterns for Small Domains with N Disulfide Bonds

N	*families*					*representative domains (95%)*				
	families with N bonds	*symmetric patterns*		*reducible patterns*		*domains with N bonds*	*symmetric patterns*		*reducible patterns*	
		O	*P*	*O*	*P*		*O*	*P*	*O*	*P*
2	37	37	37.0	1	12.3	266	266	266.0	5	88.7
3	37	21	17.3	6	12.3	506	294	236.1	130	168.7
4	10	2	2.4	3	3.0	84	2	20.0	64	24.8
5	5	0	0.4	0	1.3	11	0	0.9	0	2.8
6	1	0	3.2e-2	0	0.2	2	0	6.4e-2	0	0.4
9	1	0	7.8e-4	1	0.1	1	0	7.8e-4	1	0.1
S		**60**	**57.1**	**11**	**29.2**		**562**	**523.1**	**200**	**285.5**

Table 3.4: Bonding Patterns for Small Domains with N Disulfide Bonds. The number of observed (O) and predicted (P) families or representative domains with symmetric or reducible bonding patterns are shown.

The most striking result is that the number of observed reducible patterns is much lower than what is predicted when assuming all patterns are equally probable. This clearly suggests that irreducible disulfide bonding patterns offer some kind of evolutionary advantage over reducible patterns. The most obvious explanation would be that irreducible patterns result in more stable structures than their reducible counterparts. Because a reducible pattern can by definition be divided into independent cross-linked

regions, each subpattern found within a reducible pattern could be responsible for locally stabilizing areas within a protein, but still allow for undesirable flexibility between those regions. In the case of small protein domains that lack a hydrophobic core, the higher complexity of irreducible patterns may often be essential for maintaining the protein fold.

In contrast, occurrences of symmetric patterns in the set are slightly higher than expected if all patterns are equally probable. The total number of symmetric patterns observed in this set (60 families or 562 representative domains) is 5-8% greater than the sum predicted from random. This minor overrepresentation may indicate that some kind of biological advantage is gained by symmetric bonding patterns as well. This may also be a reflection of the high frequency of symmetric fold topologies seen in proteins.

Thus, this analysis finds that symmetry was slightly overrepresented while reducibility was highly underrepresented in the disulfide bonding patterns of small protein domains. Notably, Benham and Jafri found that both symmetry and reducibility were greatly overrepresented in their dataset (Benham and Jafri 1993). There are several explanations that account for these conflicting results. One factor is the difference in sample size: Benham and Jafri's work was completed about 12 years ago when the number of proteins with confidently established disulfide bonding patterns was quite small. Additionally, Benham and Jafri's study was not limited to small protein domains. It is likely that the biological and physical forces guiding disulfide bonding patterns are different for larger proteins, which could contribute to the differences between these results. Furthermore, their analysis considered entire polypeptide chains rather than individual domains. The inclusion of multi-domain proteins would greatly increase the observed occurrences of reducible bonding patterns relative to the current survey of only single-domain representatives.

A related analysis was performed by Hartig *et al*., who examined occurrences of each specific 2- and 3-bond pattern (Hartig, Tran et al. 2005). Their observed frequencies of those patterns are very well correlated with the bonding pattern frequencies in the current set of disulfide-rich domains.

3.5 FUNCTIONS OF DISULFIDE-RICH DOMAINS

3.5.1 General domain functions

Disulfide-rich domains have been demonstrated to accomplish a wide variety of cellular roles. The roles of these domains can be divided into three functional categories: communication, structural, and enzymatic. By far the most prevalent of these three is communication. Popular functions of disulfide-rich domains in this category are hormones, growth factors, pheromones, enzyme inhibitors, ligand-binding domains of extracellular receptors, etc. A related set of functions includes tasks of an offensive (e.g. immobilizing prey by interfering with ion channel activity) or defensive (e.g. inducing cell lysis of microbial predators) nature. With the exception of the ligand-binding domains, most disulfide-rich domains with communication roles are single-domain (i.e. not subdomains of larger polypeptides). Furthermore, these domains are predominantly extracellular.

Other disulfide-rich domains are theorized to play structural roles. Most of these examples are subdomains within larger proteins, such as the PSI domain of the human Met receptor (fold group 26) which is proposed to serve as a wedge to properly orient the propeller-like and immunoglobulin domains of this protein (Kozlov, Perreault et al. 2004). There are also a few single-domain disulfide-rich proteins with structural roles, including the hinge protein (non-heme 11 kDa protein) of the cytochrome bc_1 complex (fold group 2), which is essential for complex formation (Kim and King 1983).

Additionally, there are two disulfide-rich proteins that have been demonstrated to perform enzymatic functions. These are phospholipase A2 (fold group 10) and the light chain of methylamine dehydrogenase (fold group 22).

It should be noted, however, that many disulfide-rich domains are not yet functionally characterized. In many cases, a cellular or physiological role has been

established but the molecular target is not yet identified, and then there are some domains for which the function is completely unknown.

3.5.2 Functional convergence of disulfide-rich domains

There are many examples of similar functions that are performed by a number of unrelated disulfide-rich domains. Cases of similar functions performed by domains within different families and/or fold groups are most likely examples of convergent evolution. Of course, it is also possible that some examples may reflect remote homology that cannot be established with confidence given the currently available sequence, structure, and functional data.

The most prevalent function among the domains in this classification is inhibition of the activity of many different types of ion channels. Disulfide-rich toxins have been demonstrated to block channels that conduct a variety of different ions (including Na^+, K^+, Ca^{2+}, Cl^-, or non-specific cations) with a variety of different gating mechanisms (voltage-gated, ligand-gated, or mechanosensitive). In this classification, 9 fold groups and 10 families include at least one protein that is a known or putative ion channel inhibitor. Among these, there are several examples of disulfide-rich toxins from related species found not only in different families, but also in different fold groups: sea anemone toxins with right-handed 3-helix bundle or 3-strand antiparallel β-sheet folds (fold groups 3 and 13), scorpion toxins with short α-hairpin or knottin-like folds (fold groups 1 and 23), and conotoxins with knottin-like or small, disulfide-closed loop folds (fold groups 23 and 35). Another common function is the inhibition of various serine proteases, including trypsin, chymotrypsin, elastase, plasmin, thrombin, factor Xa, factor VIIa, etc. Despite their different specificities and global folds, many of these inhibitors are believed to share a common mechanism. Comparison of the backbone angles of the inhibitory loops of serine protease inhibitors from unrelated families has shown that these regions adopt very similar conformations (Laskowski and Qasim 2000). Serine protease inhibitors are found in 8 fold groups and 10 families of this classification. Also, many

disulfide-rich domains are annotated as antimicrobial or defensin proteins. The presumed mechanism of these domains is to induce cell lysis of a microbial predator by disrupting the cell membrane, although the details of such a mechanism are unclear. Moreover, some of these proteins are thought to target specific extracellular receptors rather than interact directly with the membrane. Putative antimicrobial or defensin proteins are found in 6 fold groups and 9 families of this classification. Membrane disruption is also the suggested mechanism for a number of non-defensive proteins, such as snake venom cardiotoxins. The functions described in the preceding examples are most commonly performed by whole (i.e. single-domain) proteins. Disulfide-rich subdomains, on the other hand, are frequently involved in the binding of molecules found in abundance on or near the cell surface, such as heparin, chitin, integrins, and TGFβ superfamily members.

3.5.3 Functional divergence of disulfide-rich domains

Examples of the divergent evolution of homologous disulfide-rich proteins to various molecular or cellular functions are common as well. Often, these domains perform related functions, such as spider toxins that block various types of ion channels (Figure 3.9a), disintegrins that inhibit the function of different integrin receptors with high selectivity, or α-conotoxins that bind to and inhibit assorted subtypes of nicotinic acetylcholine receptors. In these rapidly evolving families, numerous highly similar proteins are frequently found in the same species.

In other cases, the homologous proteins perform more distant functions. In the BPTI-like family, for example, some members inhibit serine proteases while others block K^+ or Ca^{2+} channels. The BPTI-like family also includes an interesting example of mechanistic divergence while cellular function is retained. As previously mentioned, many serine protease inhibitors appear to share a common mechanism. In these canonical inhibitors, including the majority of inhibitors in this family, the inhibitory loop forms one β-strand of a distorted antiparallel β-sheet at the active site of the protease (Laskowski and Qasim 2000). However, in a small number of BPTI-like family

Figure 3.9: Functional Divergence of Disulfide-Rich Homologs

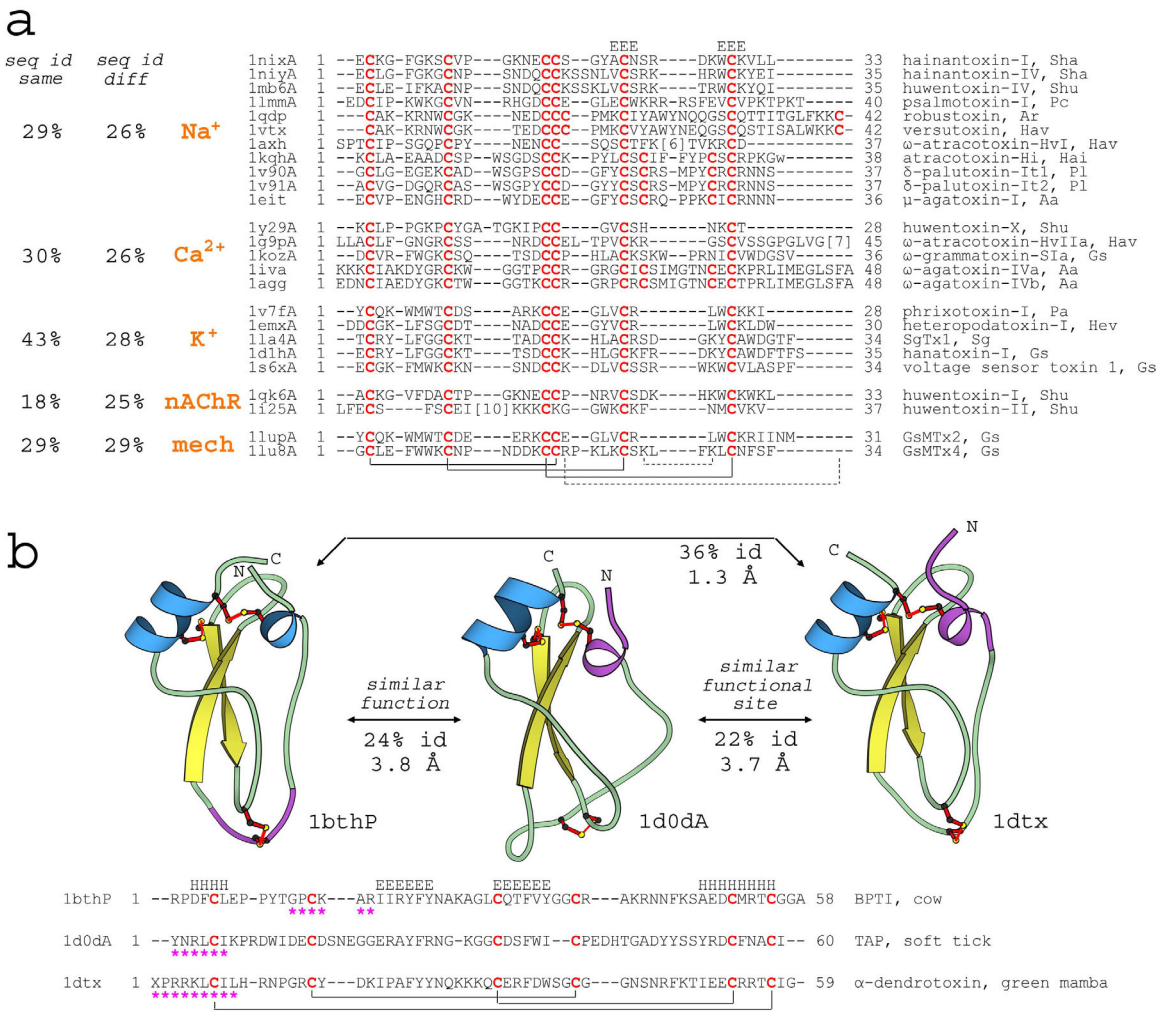


Figure 3.9: Functional Divergence of Disulfide-Rich Homologs. a) Alignment of spider toxins, grouped by channel type targeted: sodium (Na^+), calcium (Ca^{2+}), potassium (K^+), nicotinic acetylcholine receptors (nAChR), mechanosensitive (mech). "Seq id same" refers to the average sequence identity among spider toxins that block the same type of channels. "Seq id diff" refers to the average sequence identity between spider toxins that block a specific channel type and the rest of the spider toxins (inhibitors of other channel types). Percent identities were calculated using a non-redundant set of spider toxins with solved structures and known channel types. Species abbreviations are as follows: Aa *Agelenopsis aperta*, Ar *Atrax robustus*, Gs *Grammostola spatulata*, Hav *Hadronyche versuta*, Hai *Hadronyche infensa*, Hev *Heteropodidae venatoria*, Pc *Psalmopoeus cambridgei*, Pl *Paracoelotes luctuosus*, Pa *Phrixotrichus auratus*, Sha *Selenocosmia hainana*, Shu *Selenocosmia huwena*, Sg *Scodra griseipes*. Highly conserved disulfide bonds are indicated by solid lines; additional disulfide bonds in some family members are indicated by dashed lines. b) Serine protease inhibitors in the BPTI-like family perform the same function using different parts of the protein fold: BPTI (pdb|1bth, chain P) and TAP (pdb|1d0d, chain A). Other family members utilize the same region of the protein to perform different functions: the N-terminal residues are the key functional residues of TAP and green mamba snake α-dendrotoxin (pdb|1dtx). Functional sites are indicated in purple in the MOLSCRIPT diagrams and with asterisks in the alignment.

members, such as tick anticoagulant protein (TAP) and ornithodorin, the inhibitory activity is accomplished by the N-terminal residues which run parallel to the protease active site (van de Locht, Stubbs et al. 1996; St Charles, Padmanabhan et al. 2000). Notably, the N-terminal region of this structure also contributes the key functional residues of α-dendrotoxin (Gasparini, Danse et al. 1998), a non-protease-inhibitor member of this family. Interestingly, the toxin members of this family share higher sequence and structural similarity with the canonical-type inhibitors rather than the TAP-like inhibitors with which they share a common functional site (Figure 3.9b).

3.6 CONCLUSIONS

A comprehensive structural classification of small, disulfide-rich protein domains has been carried out. Nearly 3000 disulfide-rich domains were identified in the PDB and have been arranged into 41 fold groups based on structural similarity. These fold groups describe more broad structural relationships than existing groupings of these domains and therefore bring together representatives with previously unacknowledged similarities. Within the fold groups, the domains are assembled into families of homologs. 98 families of disulfide-rich domains, some of which unite previously unlinked proteins, are presented in this structural classification. This classification made possible the examination of cases of convergent and divergent evolution of functions performed by disulfide-rich proteins. Furthermore, disulfide bonding patterns in these domains were evaluated. This classification contributes to the understanding of the evolution of the protein folds and functions of disulfide-rich domains.

CHAPTER 4:
Automated assignment of protein structures to evolutionary superfamilies

4.1 INTRODUCTION

4.1.1 Background

Several structural classification schemes have been developed for the purpose of cataloguing all available protein structures, such as SCOP (Murzin, Brenner et al. 1995), CATH (Orengo, Michie et al. 1997), and Dali Domain Dictionary (Dietmann and Holm 2001). These databases are commonly used for studying structural and evolutionary relationships between proteins. Detecting remote homology between protein structures is a difficult task because of the challenge in differentiating between distant homologs and structural analogs. Several researchers have reported the inadequacy of various structural similarity measures for distinguishing homologous and analogous relationships (Russell and Barton 1994; Holm and Sander 1997; Russell, Saqi et al. 1997; Matsuo and Bryant 1999). Therefore, although the databases mentioned above are associated with automatic methods for identifying potential structural neighbors of a new protein query, they are often incapable of assigning domains to a unique position in the classification according to evolutionary relationships. Determining appropriate evolutionary relationships within a database is usually accomplished by expert manual analysis. Although manual classification of protein structures remains the gold standard, the necessity for reliable automatic tools that can reproduce the results of such a classification scheme becomes increasingly apparent as available databases continue to grow in size. Such tools must be capable of detecting homology between distantly related proteins while keeping false positives at a minimum.

Available tools for assigning proteins to existing classification schemes use either structure-based or sequence-based comparison methods. Classification predictions from

structure comparison tools like SSM (Krissinel and Henrick 2003), GRATH (Harrison, Pearl et al. 2003), and F2CS (Getz, Vendruscolo et al. 2002) are generally accurate to the fold or topology level but do not necessarily have evolutionary implications. Consequently, establishing homology between the query and the predicted neighbors often requires a more thorough examination. Classification assignments from sequence comparison tools such as SUPERFAMILY (Gough, Karplus et al. 2001) can detect homology but often miss the more remote homologous relationships suggested by structural similarities. These tools are generally reliable for homology detection in easy to moderate cases but frequently produce many false positive results for more distant relationships. A strategy combining information from both sequence and structure comparisons would be expected to perform better than either method alone by exploiting the advantages of each approach.

4.1.2 Objectives

An algorithm has been developed to map domains within protein structures with their homologs in an existing classification scheme. The general strategy employed by this algorithm is to combine the results of several existing sequence and structure comparison tools in order to determine classification assignments. The comparison tools incorporated in the algorithm each utilize a different methodology for identifying similarities between proteins, and consequently, these tools have different advantages and limitations. An approach combining different methods of homology detection is expected to capitalize on the proficiencies of each comparison tool while the limitations of those tools are neutralized by the inclusion of other methods.

This algorithm, named SCOPmap, has been developed to map domains in protein structures to the SCOP database, which is a manually curated hierarchical classification scheme based on the structural and evolutionary relationships between proteins. SCOPmap assigns protein domains at the superfamily level, which is the broadest level of homology in the SCOP database. SCOPmap also performs assignments at the SCOP fold

level when confident superfamily level assignments cannot be made. The primary application of SCOPmap is to identify domains within newly solved protein structures and assign these domains to the appropriate SCOP superfamily. The strategy employed by this algorithm is not limited to SCOP and could be applied to any other similar database or classification scheme as well.

The performance of SCOPmap has been evaluated on two test sets, each of which includes over 4500 protein domains. Comparison of SCOPmap results and SUPERFAMILY (Gough, Karplus et al. 2001) results for the same test set indicates that SCOPmap performs better than SUPERFAMILY both in terms of overall correct assignments and in accurate definition of the domain boundaries of those assignments. SCOPmap's performance at both the SCOP superfamily and SCOP fold levels has been analyzed, and the performance of the individual comparison tools incorporated in the algorithm has been evaluated. Furthermore, examples of difficult cases that are successfully mapped are described, and the reasons why some domains are not mapped by this algorithm are investigated.

4.2 METHODS

4.2.1 Mapping strategy of the SCOPmap algorithm

General Strategy

The purpose of SCOPmap is to assign domains within protein structures to the SCOP classification at the broadest level of homology, i.e. the SCOP superfamily level. The general strategy is to combine the results of several existing sequence and structure comparison tools to determine superfamily assignments as well as domain boundaries. Because the basis for identifying relationships between proteins varies between the different comparison tools, this combinatorial approach is expected to perform better than a single comparison tool alone. Furthermore, an approach utilizing multiple comparison

tools is consistent with the conclusions reached by Novotny *et al*. from an analysis of several fold comparison servers (Novotny, Madsen et al. 2004).

There are three main steps in this mapping strategy. First, hits are identified between the query protein and proteins with known SCOP assignments using several different comparison tools. Next, the results of those comparison tools are used to determine the appropriate SCOP superfamily level assignment for domains within the query. Assignments are made by a consensus-like method in which more reliable comparison tools are given preference. Finally, the algorithm uses the results of the comparison tools to define the boundaries of the domain assignments by identifying the longest non-overlapping segments.

Library of representative SCOP domains

A subset of SCOP domains with less than 40% identity to each other was downloaded from the ASTRAL database (Brenner, Koehl et al. 2000). This set contains domains from the "All alpha proteins", "All beta proteins", "Alpha and beta proteins (α/β)", "Alpha and beta proteins (α+β)", "Multi-domain proteins (alpha and beta)", "Membrane and cell surface proteins and peptides", and "Small proteins" classes of SCOP. Domains from the "Coiled coil proteins" class were manually added to the library. Results using two different SCOP libraries are discussed. One library is based on SCOP v1.61 and contains 4813 domains from 1110 SCOP superfamilies, while the other library is based on SCOP v1.63 and contains 5265 domains from 1232 superfamilies. Each library includes at least one representative of each SCOP superfamily that is present in that version of the SCOP classification.

Set of representative query chains

Input for SCOPmap is a list of PDB (Berman, Westbrook et al. 2000) identifiers. Each chain in these structures is considered as a separate query. The BLASTCLUST program (I. Dondoshansky and Y. Wolf, unpublished; ftp://ftp.ncbi.nih.gov/blast/) is used for preliminary clustering of all chains at 95% sequence identity and 95% length

coverage. A representative set of query chains is constructed from the first member of each BLASTCLUST cluster, excluding chains less than 20 residues in length. Chains less than 20 residues in length are designated as fragments and are ignored by SCOPmap.

4.2.2 Mapping step 1: Identifying hits between query and library domains using existing comparison methods

The gapped BLAST (Altschul, Madden et al. 1997), RPS-BLAST(Marchler-Bauer, Anderson et al. 2003), PSI-BLAST (Altschul, Madden et al. 1997), COMPASS (Sadreyev and Grishin 2003), MAMMOTH (Ortiz, Strauss et al. 2002), and DaliLite (Holm and Park 2000) tools are used in SCOPmap. The first four of these are sequence comparison tools and are listed in order of increasing sensitivity to remote homologs: query sequence against a database of sequences (gapped BLAST), query sequence against a database of profiles (RPS-BLAST), query profile against a database of sequences (PSI-BLAST), and query profile against a database of profiles (COMPASS). The two structure comparison tools used are the MAMMOTH and DaliLite algorithms. Furthermore, in an effort to improve performance on the detection and assignment of small protein domains, SCOPmap assesses the ratio of DaliLite scores for small domains in non-self versus self comparisons. Additionally, SCOPmap includes two tools which incorporate elements of both sequence and structure comparisons: correlation of conservation patterns in structurally aligned regions, and the agreement of pairwise alignments produced by structure comparison tools (DaliLite or MAMMOTH) with those produced by sequence comparison tools (gapped BLAST, RPS-BLAST, or PSI-BLAST). Thus, eight different comparison methods are used to identify similarities between query and library proteins. Each of these eight methods is described in detail below.

Method 1) gapped BLAST: query sequence against database of sequences

Gapped BLAST (Altschul, Madden et al. 1997) is run for each representative query sequence against sequences of all chains from PDB structures in SCOP (37,007

sequences in SCOP v1.61; 41,066 sequences in SCOP v1.63). The criteria for an accepted BLAST hit are an E-value $\leq$ 0.005 and coverage of all but 10 residues at each end of both the query and database sequences. Hits are also accepted if the query and library sequences are at least 80% identical and all but 10 residues at each end of the query sequence are covered by the alignment, irrespective of E-value. Because the database sequences used for gapped BLAST are complete chains, the accepted hits are then converted from library chains to library domains according to the SCOP-defined domain boundaries of those library sequences. This conversion is not necessary for accepted hits from the other seven comparison methods since the library representatives in those methods are domains rather than complete chains. For all queries for which the entire length of the chain (except for 10 residues at each termini) corresponds to an accepted BLAST hit, superfamily assignments are based solely on the BLAST results and no other comparison tools are used. All query chains with no BLAST hits passing the described criteria are submitted to each of the remaining methods.

Method 2) RPS-BLAST: query sequence against database of profiles

RPS-BLAST (Marchler-Bauer, Anderson et al. 2003) is run for the query sequence against a database of profiles for the library of representative SCOP domains. Profiles were constructed for each library domain by running PSI-BLAST against the non-redundant database for 5 iterations or until convergence with an E-value cutoff of 0.005. The criteria for an accepted RPS-BLAST hit are an E-value $\leq$ 0.005 and coverage of all but 10 residues at each end of the library domain.

Method 3) PSI-BLAST: query profile against database of sequences

A profile for the query sequence is constructed by running PSI-BLAST (Altschul, Madden et al. 1997) against the non-redundant protein database for 5 iterations or until convergence with an E-value cutoff of 0.001. This profile is subsequently used as input for a PSI-BLAST search against a database of all SCOP domain sequences (42,465 domain sequences in SCOP v1.61; 47,013 domain sequences in SCOP v1.63). The

criteria for an accepted PSI-BLAST hit are an E-value $\leq 10^{-4}$ and coverage of all but 10 residues at each end of the SCOP domain database sequence.

Method 4) COMPASS: query profile against database of profiles

The profiles for the query (constructed in the PSI-BLAST step) and the SCOP library domains (constructed in the RPS-BLAST step) are prepared for COMPASS (Sadreyev and Grishin 2003) by: 1) deleting all columns with gaps in the query sequence, 2) removing all sequences identical to the query, and 3) retaining only 1 copy of any sequences in the profile that have greater than 97% identity. COMPASS is then run for the query profile against each of the SCOP library domain profiles. Accepted COMPASS hits have an E-value $\leq 10^{-10}$ and coverage of all but 10 residues at each end of the library domain.

Method 5) MAMMOTH: query structure against database of structures

The query structure is compared to each library domain structure via MAMMOTH (Ortiz, Strauss et al. 2002). For each query-library domain pair, the MAMMOTH Z-score (Z_M) and the normalized BLOSUM (Henikoff and Henikoff 1992) score for the pairwise alignment made by MAMMOTH (BS_M) are calculated. MAMMOTH hits are accepted if they meet all of the following criteria:

1) $Z_M \geq 4.0$;

2) coverage of $\geq$50% of the library domain;

3) ($BS_M \geq 0.75/Z_M + 0.1$) or ($Z_M \geq 22.0$).

The cutoffs for accepted hits were determined based on the MAMMOTH Z-score (Z_M) and BLOSUM score (BS_M) of 106,310 randomly chosen pairs of SCOP domains from SCOP v1.61 (Figure 4.1). Approximately 1/3 of these pairs of domains belong to the same SCOP superfamily while the remaining 2/3 of the pairs belong to different SCOP superfamilies.

Figure 4.1: Threshold for Accepting MAMMOTH Hits

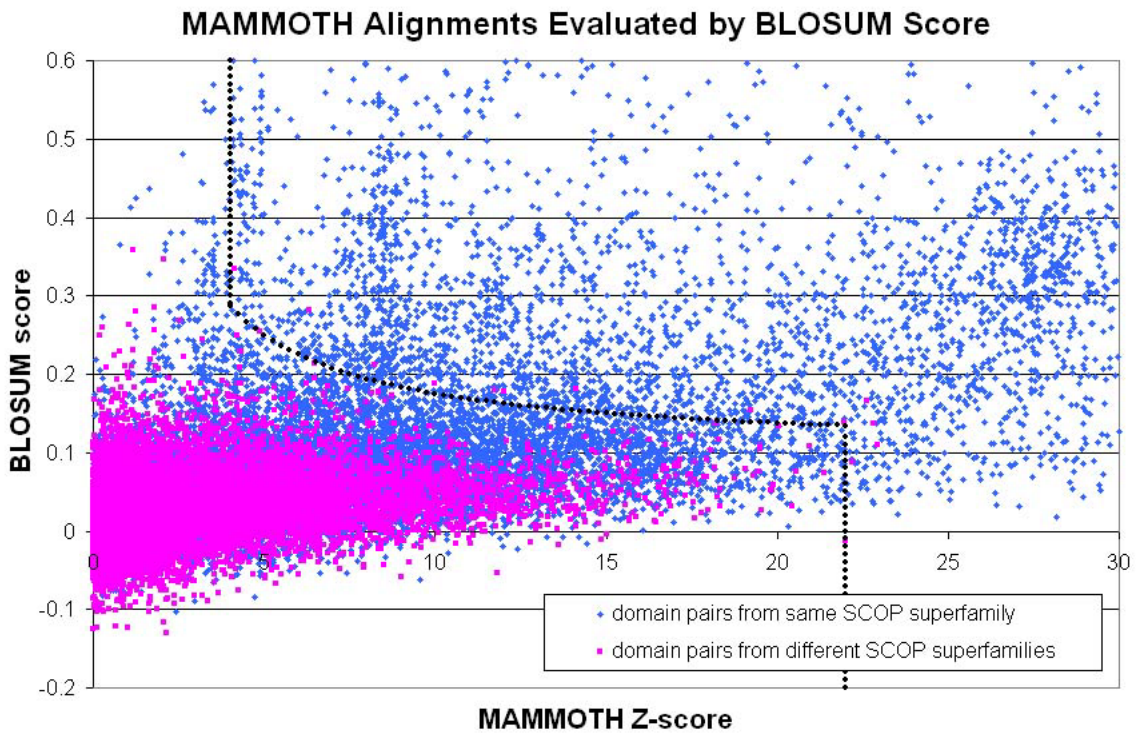


Figure 4.1: Threshold for Accepting MAMMOTH Hits. Data are MAMMOTH results for >100,000 randomly chosen pairs of SCOP domains. Black dotted lines indicate chosen thresholds for accepting a MAMMOTH hit.

Method 6) DaliLite: query structure against library structure comparisons

Additional structure comparisons are performed with DaliLite (Holm and Park 2000) for queries with a segment of 20 residues or longer that did not correspond to an accepted MAMMOTH hit. Query-library domain pairs for which $BS_M \geq -0.01*Z_M - 0.14$, $Z_M > 0$, and the pairwise alignment made by MAMMOTH covered at least 40% of the library domain are identified. The score cutoffs for selecting pairs for comparison via DaliLite were determined by evaluating the MAMMOTH Z-scores (Z_M) and BLOSUM scores (BS_M) for randomly chosen pairs of SCOP domains that pass the DaliLite score cutoffs (see below) but fail the MAMMOTH score cutoffs. The threshold was chosen by determining the score cutoffs that would identify the most number of pairs passing the

DaliLite cutoffs and the fewest pairs failing the DaliLite cutoffs, thereby maximizing the number of potential accepted hits while minimizing the overall computation time required (Figure 4.2). If more than 200 query-library domain pairs meet these criteria, only the 200 pairs with the highest Z_M scores are selected. If no pairs meet these criteria, the 50 query-library domain pairs with the highest Z_M scores are selected.

Figure 4.2: Selecting Domains Pairs for Submission to DaliLite

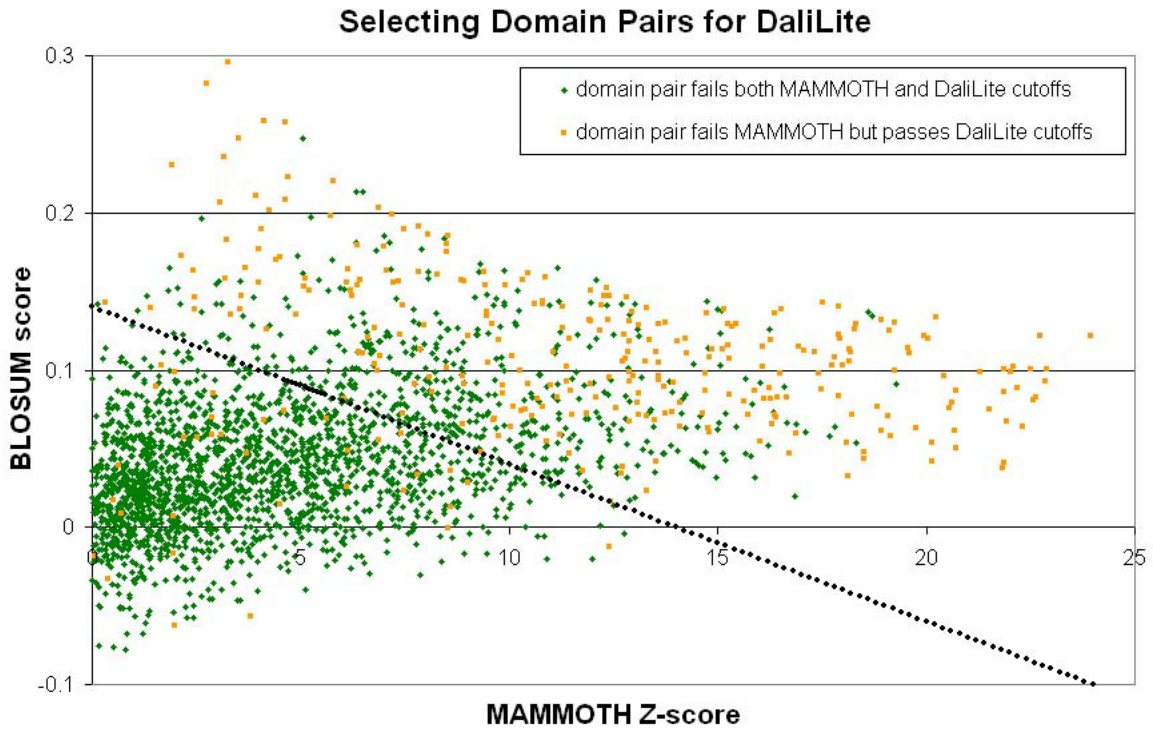


Figure 4.2: Selecting Domain Pairs for Submission to DaliLite. Data are MAMMOTH results for 2500 randomly chosen pairs of SCOP domains that fail the MAMMOTH acceptance criteria. The black dotted line indicates chosen threshold for accepting a pair for submission to DaliLite.

DaliLite structure comparison is performed for each of the selected query-library domain pairs, and the DaliLite Z-score (Z_D) and the normalized BLOSUM score for the pairwise alignment made by DaliLite (BS_D) are calculated. Hits are accepted if they meet one of the following sets of criteria:

1) $Z_D \geq 4.0$, $BS_D \geq -0.01*Z_D + 0.27$, and coverage of $\geq$50% of the library domain;

2) $BS_D \geq 0.4$ and coverage of $\geq$50% of the library domain;

3) $Z_D \geq 14.0$ and coverage of $\geq$50% of the library domain.

The cutoffs for accepted hits were determined based on the DaliLite Z-score (Z_D) and BLOSUM score (BS_D) of 4000 randomly chosen pairs of SCOP domains from SCOP v1.61, with half of these pairs belonging to the same superfamily and half of the pairs belonging to different superfamilies (Figure 4.3).

Figure 4.3: Threshold for Accepting DaliLite Hits

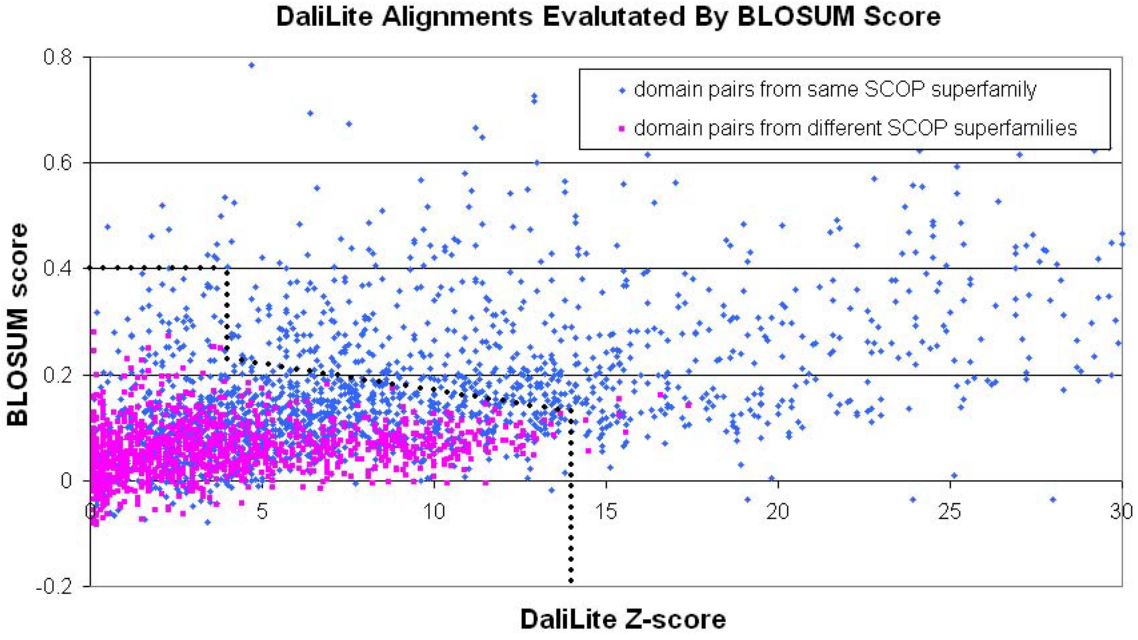


Figure 4.3: Threshold for Accepting DaliLite Hits. Data are DaliLite results for 4000 randomly chosen pairs of SCOP domains. Black dotted lines indicate chosen threshold for accepting a DaliLite hit.

As discussed in Chapter 3, automated methods often perform unreliably for small proteins. Furthermore, the E-value and Z-score thresholds used by SCOPmap were determined based on comparisons within SCOP v1.61 and are consequently heavily biased in favor of "normal" proteins (>100 residues); in v1.61, less than 8% of superfamilies belong to the "Small proteins" class of SCOP. In an effort to evaluate

small protein domains more aptly, SCOPmap assesses the ratio of structure comparison scores for non-self vs self-comparisons. For any DaliLite comparison of a query-library domain pair in which the library domain is less than 150 residues in length, the Z-score ratio is calculated by $Z_{ratio} = (D_{Z, Q\text{-}L})/(D_{Z, L\text{-}L})$, where $D_{Z, Q\text{-}L}$ is the DaliLite Z-score for the query-library domain pair and $D_{Z, L\text{-}L}$ is the DaliLite Z-score for the library representative against itself. A hit is accepted if both $Z_{ratio} > 0.3$ and $D_{Z, Q\text{-}L} > 4$ are satisfied. These cutoffs were determined based on DaliLite Z-scores from 2000 randomly chosen pairs of SCOP domains from SCOP v1.61, where at least one domain within a pair is less than 150 residues in length (Figure 4.4).

Figure 4.4: Threshold for DaliLite Z-score Ratios

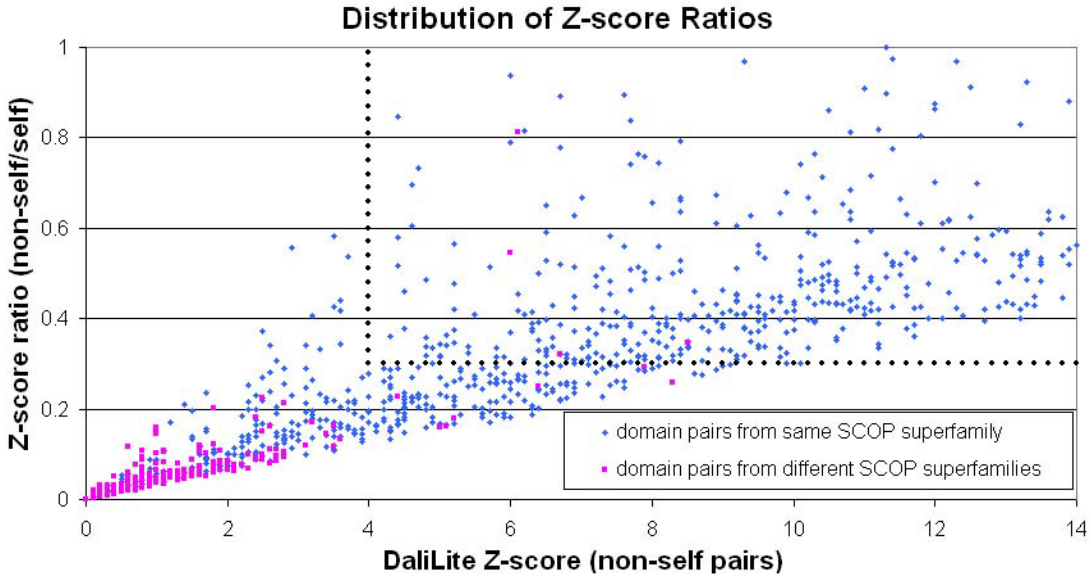


Figure 4.4: Threshold for DaliLite Z-score Ratios. Data are calculated from DaliLite results from 1000 randomly chosen pairs of SCOP domains. Black dotted lines indicate chosen thresholds.

Method 7) CSV: correlation of conservation patterns

Because homologous domains often have similar conservation patterns, the degree of correlation between the conservation patterns of two domains can be used for remote homolog detection. Distant homologs typically display drastically diminished

overall sequence similarity. Thus, such cases of remote homology are more likely to be identified by conservation pattern analysis, which considers only the most conserved residues, rather than by typical sequence comparison methods, which are highly dependent on overall sequence similarity. Conservation scores for query-library domain pairs are calculated by two methods: using a conservation substitution matrix and using the COMPASS algorithm.

The query-library domain pairs selected for conservation pattern comparison are determined based on the results of the DaliLite pairwise comparisons in the previous method. The correlation of conservation patterns are calculated for all query-library domain pairs with $Z_D \geq 4.0$, or for the 20 pairs with highest DaliLite Z-score ($Z_D \geq 2.0$ required) if no pairs have DaliLite Z-score greater than 4. Only pairs for which the library domain profile (constructed for the RPS-BLAST step and modified for the COMPASS step) contains 5 or more sequences are considered. The AL2CO algorithm (Pei and Grishin 2001) (window size 3) is used to calculate the entropy-based conservation index for each position in the query profile and in the library domain profile. DaliLite-aligned positions scoring in the top 25% of either profile are selected (henceforth referred to as the chosen positions).

Any two given positions from the profiles of the query and library domains can be compared to determine their similarity in terms of conservation patterns. The degree of correlation between those conservation patterns is referred to as the position-pair conservation score. For example, if both positions are highly conserved, the position-pair conservation score for that specific pair will be high. Conversely, if one position is highly conserved while the amino acid distribution in the other position is random, the position-pair conservation score will be low. In the first scoring system, position-pair conservation scores are determined based on the entropy-based conservation indices for the chosen positions with a conservation substitution matrix used as a scoring matrix. Then, the scoring matrix-based conservation score is calculated for the query-library domain pair by:

$$CSV_{cons,D} = [S_n - S_{rand}]/[(S_1+S_2)/2 - S_{rand}],$$

where S_n is the sum of position-pair conservation scores of the aligned query positions vs. library domain positions ("chosen positions" only, see above), S_1 is the sum of position-pair conservation scores of the chosen query positions against themselves (query positions vs. query positions), S_2 is the sum of position-pair conservation scores of the chosen library domain positions against themselves (library domain positions vs. library domain positions), and S_{rand} is the sum of position-pair conservation scores of the chosen positions for all-against-all query positions vs. library domain positions normalized over length.

A COMPASS-based conservation score is also calculated for each query-library domain pair. In this scoring system, a COMPASS-based position-pair score, which describes the similarity between any two given positions, is determined based on the methodology introduced in the COMPASS method (Sadreyev and Grishin 2003). Then, the COMPASS-based conservation score for the query-library domain pair is calculated by:

$$CSV_{compass,D} = [CS_n - CS_{rand}]/[(CS_1+CS_2)/2 - CS_{rand}],$$

where CS_n is the sum of COMPASS-based position-pair scores of the aligned query positions vs. library domain positions ("chosen positions" only, see above), CS_1 is the sum of COMPASS-based position-pair scores of the chosen query positions against themselves (query positions vs. query positions), CS_2 is the sum of COMPASS-based position-pair scores of the chosen library domain positions against themselves (library domain positions vs. library domain positions), and CS_{rand} is the sum of COMPASS-based position-pair scores of the chosen positions for all-against-all query positions vs. library domain positions normalized over length.

Conservation score hits are accepted if they meet one of the following sets of criteria:

1) $CSV_{cons,D} \geq 0.1$ and $Z_D \geq 5$;

2) $CSV_{cons,D} \geq 0.25$ and $Z_D \geq 2$;

3) $CSV_{compass,D} \geq 0.4$ and $Z_D \geq 5$;

4) $CSV_{compass,D} \geq 0.5$ and $Z_D \geq 2$.

These cutoffs for accepting hits were determined based on the $CSV_{cons,D}$ scores, $CSV_{compass,D}$ scores, and DaliLite Z-scores of 4000 randomly chosen pairs of SCOP domains from SCOP v1.61 (Figure 4.5).

Figure 4.5: Threshold for Conservation Scores in DaliLite Hits

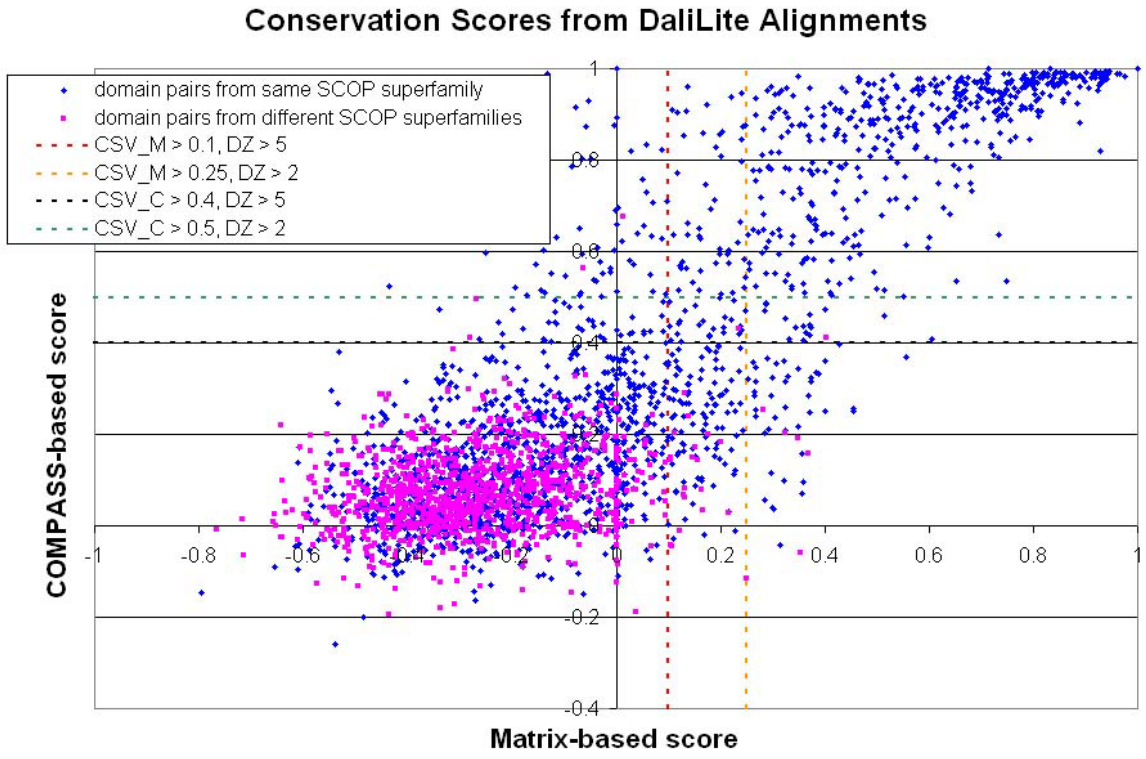


Figure 4.5: Threshold for Conservation Scores in DaliLite Hits. Data are DaliLite results for 4000 randomly chosen pairs of SCOP domains. Dashed lines indicate chosen threshold for accepting a conservation pattern hits from DaliLite alignments. Abbreviations are as follows: matrix-based conservation pattern score (CSV_M), COMPASS-based conservation pattern score (CSV_C), DaliLite Z-score (DZ).

In cases for which the DaliLite program produces no output, conservation pattern analysis is performed using pairwise alignment produced by MAMMOTH instead of FSSP alignments. The conservation analysis is done for the query-library domain pairs that would have otherwise been submitted to the DaliLite algorithm for structural comparison. Only those residue pairs in which the C_α atoms are located within 4 Å, which are indicated by an asterisk (*) by the MAMMOTH algorithm, are considered.

Again, a window size of 3 is used in the AL2CO program and only the top scoring 25% of positions are used for calculating the conservation scores. Matrix-based and COMPASS-based conservation scores are calculated as described above. Conservation score hits based on MAMMOTH alignments are accepted if they meet one of the following sets of criteria:

1) $CSV_{cons,M} \geq 0.3$ and $Z_M \geq 4$;

2) $CSV_{compass,M} \geq 0.4$ and $Z_M \geq 4$

These cutoffs for accepting hits were determined based on the $CSV_{cons,M}$ scores, $CSV_{compass,M}$ scores, and MAMMOTH Z-scores of 2000 randomly chosen pairs of SCOP domains from SCOP v1.61 (Figure 4.6).

Figure 4.6: Threshold for Conservation Scores in MAMMOTH Hits

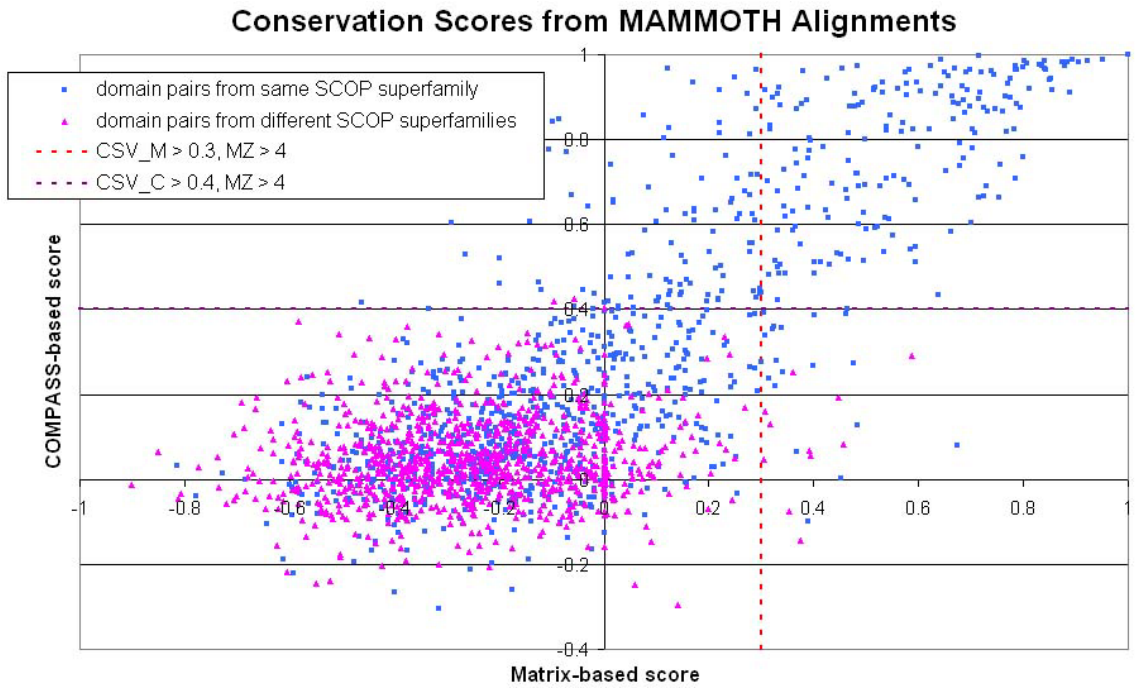


Figure 4.6: Threshold for Conservation Scores in MAMMOTH Hits. Data are MAMMOTH results for 2000 randomly chosen pairs of SCOP domains. Dashed lines indicate chosen threshold for accepting a conservation pattern hits from MAMMOTH alignments. Abbreviations are as follows: matrix-based conservation pattern score (CSV_M), COMPASS-based conservation pattern score (CSV_C), MAMMOTH Z-score (MZ).

Method 8) agreement of DaliLite or MAMMOTH alignments with gapped BLAST, RPS-BLAST, or PSI-BLAST alignments

Remote evolutionary links between protein domains can be gleaned using a combination of sequence and structural information, even when neither of these methods alone is capable of providing convincing evidence for common descent. In other words, confidence can be gained in marginal hits from one comparison tool by identifying corroborating results from another comparison tool. In this method, the degree of correlation between a pairwise alignment made by structural methods and alignments made by the sequence comparison methods is determined so that DaliLite or MAMMOTH can be used to evaluate potential hits from BLAST, RPS-BLAST, or PSI-BLAST. For any query-library domain pair with $Z_D > 0$ and BLAST, PSI-BLAST, or RPS-BLAST E-value ≤ 100, the number of correctly aligned residues (N_{ali}) in the sequence alignment is calculated using the DaliLite alignment as a reference. Hits are accepted for which $Z_D > 0$, E-value ≤ 100, and $N_{ali} \geq 15$. These cutoffs were determined based on the DaliLite Z-scores, E-values, and number of equivalently aligned residues from 1000 randomly chosen pairs of SCOP domains from SCOP v1.61 (Figure 4.7).

Figure 4.7: Threshold for Agreement of DaliLite and BLAST Alignments

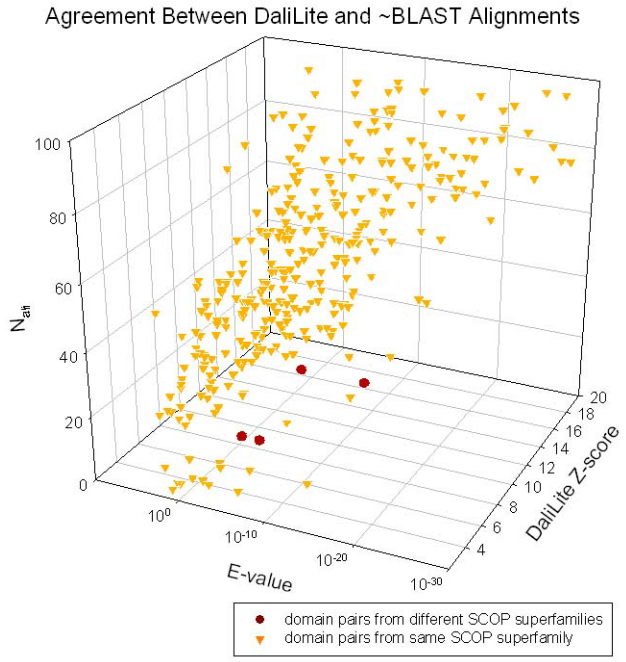


Figure 4.7: Threshold for Agreement of DaliLite and BLAST Alignments. Data are DaliLite Z-scores and gapped BLAST, RPS-BLAST, or PSI-BLAST E-values for 1000 randomly chosen pairs of SCOP domains, although very few pairs of domains from different SCOP superfamilies result in generated output from both DaliLite and the BLAST algorithms. N_{ali} refers to the number of equivalently aligned positions between the two methods (DaliLite and one BLAST method).

If an error occurs while running DaliLite for the query domain, agreement of the MAMMOTH alignment and BLAST, RPS-BLAST, or PSI-BLAST alignments is instead calculated for the same potential hits. In these cases, hits are accepted for which $Z_M >$ 2.0, E-value $\leq$ 100, and $N_{ali} \geq 15$. These cutoffs were determined based on the MAMMOTH Z-scores, E-values, and number of equivalently aligned residues from 1000 randomly chosen pairs of SCOP domains from SCOP v1.61 (Figure 4.8).

Figure 4.8: Threshold for Agreement of MAMMOTH and BLAST Alignments

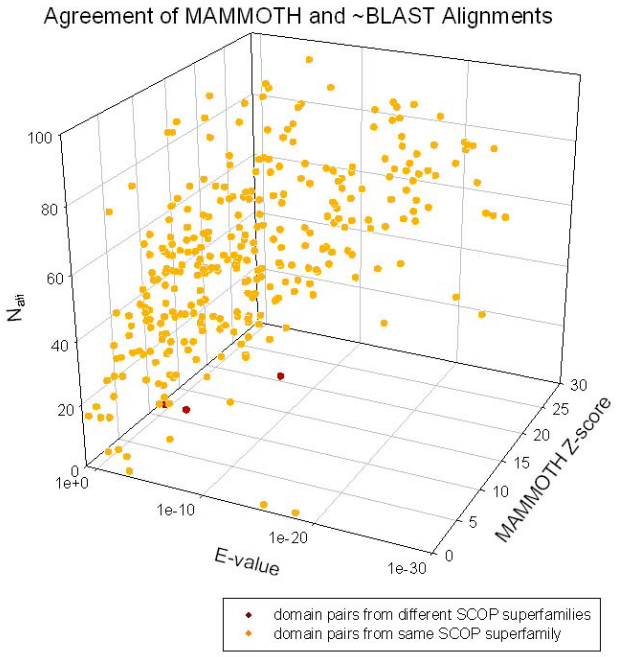


Figure 4.8: Threshold for Agreement of MAMMOTH and BLAST Alignments. Data are MAMMOTH Z-scores and gapped BLAST, RPS-BLAST, or PSI-BLAST E-values for 1000 randomly chosen pairs of SCOP domains, although very few pairs of domains from different SCOP superfamilies results in generated output from both MAMMOTH and the BLAST algorithms. N_{ali} refers to the number of equivalently aligned positions between the two methods (MAMMOTH and one BLAST method).

4.2.3 Mapping step 2: Assigning domains from query chains to SCOP superfamilies

Accepted hits from the sequence and structure comparison methods are mapped onto the query chain and domains within the chain are then assigned to SCOP superfamilies. In cases where accepted hits from multiple SCOP superfamilies mapped to the same region of the query chain, SCOPmap attempts to choose only one correct

SCOP superfamily assignment. If the overlap between two different SCOP superfamily representatives covers <50% of both domains, the conflict is resolved by the domain boundary definition (see "Mapping Step 3" below). Otherwise, SCOPmap attempts to determine which SCOP superfamily among the accepted hits is most likely to be the correct assignment.

First, for each of two conflicting assignments, all accepted hits that overlap by at least 75% and are from the same SCOP superfamily are identified. For each set of accepted hits (one set corresponding to each of the conflicting SCOP superfamilies), the number of methods that identified accepted hits to that SCOP superfamily is determined. If one SCOP superfamily is found by more methods than the other SCOP superfamily, the assignment with hits from the greater number of methods is accepted as correct. If both SCOP superfamilies are identified by an equal number of methods, the priority of those methods is used to choose the correct SCOP superfamily. The methods are ranked by reliability, which was subjectively determined based primarily on the observed number of false positives accepted by a given method during SCOPmap development. Priority rankings are as follows: BLAST > RPS-BLAST or PSI-BLAST > MAMMOTH or DaliLite > COMPASS > conservation pattern correlation or agreement of DaliLite and sequence method alignments or Z-score ratio. If both SCOP superfamilies are found by methods with equivalent priorities, the Z-scores and E-values of the hits are evaluated. If only one of the two conflicting SCOP superfamilies has E-values from any sequence comparison method below 10^{-10} or Z-scores (Z_M or Z_D) above 14.0, that SCOP superfamily assignment is accepted as correct. If a SCOP superfamily assignment has still not been made, the domain assignments to that query chain are flagged as unresolved. Of the 4580 tweaking set domains, only 25 domains (0.5%) were unassigned due to unresolved choice between conflicting SCOP superfamilies. The results obtained by inverting the order of these two steps (e.g. first comparing E-values and Z-scores, and then considering priority rankings of the eight methods) were also evaluated. There were no cases where the inverted order gave additional correct assignments, and there was a small number of cases that could be resolved by the original strategy but not by the

inverted strategy. Thus, the methodology described above is used for choosing between conflicting superfamily assignments.

4.2.4 Mapping step 3: Defining boundaries of domain assignments

The final step of the mapping algorithm is to determine the boundaries of the domains that were identified and assigned in the previous steps. Domain boundary definitions are assigned by identifying the longest non-overlapping domain assignments, with priority given to assignments made by structure comparison methods. First, DaliLite is run for all query-library domain pairs found by MAMMOTH, and the DaliLite range is used in place of the MAMMOTH range unless there is an error in the DaliLite output. This is done because DaliLite typically defines more accurate domains than MAMMOTH, and DaliLite is also more adept at recognizing large insertions within domains. Ranges of accepted hits are then given priority rankings based on which method determined the range of that hit. DaliLite ranges have highest priority, followed by MAMMOTH ranges, and then all ranges by any other comparison method. The longest non-overlapping segments with the highest priority rankings are then identified. A 3-residue cushion for overlap is allowed. Overlapping domains for which boundaries cannot be reconciled within 3 residues are flagged as unresolved boundary definitions and no domain assignment is made for that query. Of 4580 tweaking set domains, only 3 domains (0.1%) were unassigned due to unresolved domain boundary definition.

4.2.5 Assignments at the SCOP fold level

For query chains with a segment at least 20 residues in length which is not assigned to a SCOP superfamily, mapping at the SCOP fold level is attempted. In the SCOPmap algorithm, MAMMOTH is run comprehensively against the library of representative structures. Therefore, no additional comparisons must be made in order for fold level assignments to be determined. For this reason, MAMMOTH is used for

fold level assignments rather than DaliLite, which is typically run against less than 5% of the library domains. The single criterion for potential SCOP fold assignment is a MAMMOTH Z-score > 10. Fold level assignments are made by selecting the hit to an unmapped region with the highest MAMMOTH Z-score (>10) that also covers at least 50% of the library domain. The fold level Z-score cutoff was determined based on the MAMMOTH Z-scores of 106,310 randomly chosen pairs of SCOP domains from SCOP v1.61. These same pairs of domains were used for determining the superfamily assignment cutoffs. Approximately 2/3 of these pairs of domains belong to the same SCOP fold while the remaining 1/3 of the pairs belong to different SCOP folds.

4.2.6 Description of test sets

SCOPmap performance was evaluated on two separate test sets. The first set is comprised of the proteins that are included in SCOP v1.63 but not in SCOP v1.61. SCOPmap was run using a library based on the previous SCOP release (v1.61), and the SCOPmap domain assignments were compared to the SCOP-defined classification in subsequent SCOP release (v1.63). This set contains 5133 SCOP-defined protein domains, but analysis of SCOPmap performance is based only on the 4580 SCOP-defined domains with evolutionary relevance: 464 low resolution structure domains, 63 peptides, 21 designed proteins, and 5 domains that were later removed from the database are intentionally excluded. The first test set was used to establish whether the score cutoffs for the individual comparison tools used by SCOPmap were strict enough to avoid false positive assignments. After first running SCOPmap for this set of domains, a false positive rate of ~1.5% was observed. The score thresholds for some of the individual comparison tools were subsequently made more strict in order to avoid all false positive assignments in this set. For example, the E-value cutoff for PSI-BLAST was changed from $5x10^{-3}$ to $1x10^{-4}$, and the E-value cutoff for COMPASS was adjusted from $1x10^{-4}$ to $1x10^{-10}$. Because some of the domains in this set were considered while establishing the score thresholds, the first test set is more correctly described as a

"tweaking" set rather than a testing set. This set was also used for comparison to SUPERFAMILY, for which the score threshold was also chosen specifically for the purpose of precluding false positive assignments. The recommended 0.02 E-value cutoff for SUPERFAMILY, which would allow for the correct assignment of only an additional ~1% of the tweaking set domains, was not chosen due to the 4.3% false positive rate it incurs. Instead, the E-value cutoff was set at $1x10^{-5}$, the maximum value for which no false positive assignments were observed. For this comparison, the SUPERFAMILY algorithm was used with the library of SAM (Karplus, Barrett et al. 1998) hidden Markov models based on SCOP v1.61.

The second set of domains used to evaluate SCOPmap performance contains proteins included in SCOP v1.65 but not in SCOP v1.63. The second test set can be considered a true testing set. The testing set contains 5335 SCOP-defined protein domains, but only the 4941 SCOP-defined domains with evolutionary relevance were used for analysis of SCOPmap performance. Low resolution structures, peptides, and designed proteins were ignored. The library of SCOP representative domains used for mapping the queries in this set is based on SCOP v1.63.

4.2.7 Using SCOPmap to identify homologs between SCOP superfamilies

SCOPmap can also be used to identify potentially homologous proteins that belong to different SCOP superfamilies. Detection of such homologs is accomplished with a slightly altered strategy from the mapping algorithm described above. The modified algorithm evaluates one SCOP superfamily at a time by attempting to detect potential hits to SCOP domains belonging to other superfamilies via the comparison methods described above. A set of query domains is constructed from the domains that are currently included in that SCOP superfamily (based on SCOP v1.63). As in the original mapping algorithm, the query sequences are first clustered at high sequence identity to reduce the computational time. Next, each of the eight comparison methods described above is employed for each representative query. In the original mapping

strategy, queries for which accepted hits are detected via gapped BLAST are not submitted to any of the other comparison methods. However, in this modified strategy, all comparison tools are run for all representative queries, regardless of the results of the gapped BLAST step. The output is a list of all accepted hits from each of the comparison methods to SCOP domains that do not belong to the query superfamily. All hits to SCOP domains within the query superfamily are simply ignored and excluded from the output. Finally, manual analysis of potential hits was performed for selected examples in order to evaluate the significance of those hits and to determine whether an evolutionary link is likely to exist between the two SCOP superfamilies in question.

4.3 RESULTS

4.3.1 Evaluation of SCOPmap performance on two sets of queries

Mapping of the tweaking set domains

Results of SCOPmap performance on the tweaking set are shown in Table 4.1 (see section 4.2.6 for description of tweaking and testing sets). Correct SCOP superfamily assignments were made for 87.8% of the tweaking set domains. For an additional 0.3% of the tweaking set domains, the superfamily assigned by SCOPmap is not the same as the SCOP-assigned superfamily. However, in each of these cases, the superfamily assigned by SCOPmap and the superfamily specified by SCOP are homologous. For example, SCOPmap assigns the 7-bladed β-propeller domain of an archaeal surface layer protein to a homologous SCOP superfamily of 6-bladed β-propellers (Jing, Takagi et al. 2002). Because the purpose of the SCOPmap is to assign domains at the broadest level of homology in the classification (i.e. the SCOP superfamily level), such cases are not considered false positives but instead reflect special cases in the SCOP database. 6.2% of the tweaking set domains were given no superfamily assignment by SCOPmap, but are domains that belong to SCOP

Table 4.1: Automatic Mapping of PDB Structures to SCOP Superfamilies

Result	*v1.61-v1.63 test set*						*v1.63-v1.65 test set*	
	SCOPmap		*SCOPmap, sequence comparisons only*		*SUPERFAMILY*		*SCOPmap*	
	# of domains	*% of test set*	*# of domains*	*% of test set*	*# of domains*	*% of test set*	*# of domains*	*% of test set*
Assignment to correct SCOP superfamily, boundaries accurate within 10 residues	3730	**81.4%**	3507	**76.6%**	3211	**70.1%**	4136	**83.7%**
Assignment to correct SCOP superfamily, boundaries *not* accurate within 10 residues	292	**6.4%**	211	**4.6%**	662	**14.4%**	372	**7.5%**
Domain belongs to a *new* SCOP superfamily, no assignment made	284	**6.2%**	289	**6.3%**	241	**5.3%**	154	**3.1%**
Acceptable assignment, but not the same assignment as given in SCOP	13	**0.3%**	0	**0%**	71	**1.5%**	12	**0.2%**
Incorrect assignment	0	0%	0	0%	0	0%	7	0.2%
Domain belongs to an *existing* SCOP superfamily, no assignment made	261	5.7%	573	12.5%	395	8.6%	260	5.3%

Table 4.1: Automatic Mapping of PDB Structures to SCOP Superfamilies. Boldface text indicates correct assignments.

superfamilies that are new in v1.63. Because such domains cannot be appropriately assigned to a superfamily that is represented in the library used by SCOPmap (v1.61 in this case), these are also considered correctly mapped (i.e. true negative assignments). Thus, a total of 94.3% of the tweaking set domains are correctly mapped by SCOPmap. The remaining 5.7% of the tweaking set are false negative assignments. These domains belong to superfamilies that exist in SCOP v1.61, but no superfamily assignment is made by SCOPmap.

Mapping of the testing set domains

Results of SCOPmap performance on the testing set are also shown in Table 4.1. Correct SCOP superfamily assignments were made for 91.2% of the testing set domains.

In an additional 0.2% of the test set, the domain assignments given by SCOPmap are homologous to the superfamilies specified by SCOP. 3.1% of the tweaking set domains are given true negative assignments. These are cases in which the appropriate superfamily assignment is not a part of the library used by SCOPmap (based on SCOP v1.63 in this case), and no superfamily assignment is made by SCOPmap. Thus, a total of 94.5% of the testing set domains are given correct assignments by SCOPmap. 5.3% of the testing set domains are false negative assignments in which the domain belongs to a superfamily that is present in SCOP v1.63, but no superfamily assignment is made by SCOPmap. The remaining 0.2% of the testing set domains are given false positive assignments.

4.3.2 False positive assignments in the testing set

Because the score cutoffs used by SCOPmap's individual comparison tools were determined while considering domains from the tweaking set, those cutoffs were therefore influenced by the specific collection of domains in that set. Had a different test set been considered when establishing these cutoffs, it is likely that the score cutoffs would be slightly different. Thus, the few false positive assignments observed in the second test set are not unexpected. Furthermore, the number of false positive domain assignments made is higher than the number of incorrect hits between query and library domains that are accepted. Due to redundancy in the test set (e.g. often one structure contains several identical chains and therefore several identical domains), the 7 domains mapped incorrectly essentially reflect only 3 different examples of false positive assignments.

Each incorrectly assigned domain has less than 10% sequence identity to the nearest library representative from the same SCOP superfamily. Furthermore, all of the false positive assignments are due to scores from the individual comparison tools that barely meet the cutoffs required for acceptance. Such cases reflect the influence that a few specific domains can have in determining the exact values of the minimum score

threshold requirements. All incorrect assignments were made due to a hit accepted by one of the comparison tools that includes both sequence and structure components.

For example, addiction antidote protein MazE from *Escherichia coli* (pdb|1mvf, chains D and E (Loris, Marianovsky et al. 2003); SCOP domains: d1mvfd_ and d1mvfe_) belongs to the Kis/PemI addiction antidote superfamily in SCOP and forms a pseudobarrel as a homodimer. SCOPmap incorrectly maps this protein to the "Transcription-state regulator AbrB, the N-terminal DNA recognition domain" superfamily in SCOP, which is a 2-layer α/β protein. This assignment is due to a hit found to the N-terminal DNA recognition domain of AbrB from *Bacillus subtilis* (pdb|1ekt (Vaughn, Feher et al. 2000); SCOP domain: d1ekta_). Although the aligned regions of these two domains have the same secondary structure (an α-helix, a β-strand, and followed by a β-hairpin) and similar spatial arrangement, the overall topologies of these folds are highly dissimilar. This hit is accepted due to the 18 pairs of residues from the query and library representative which are equivalently aligned in pairwise alignments produced by PSI-BLAST (E-value = 55) and DaliLite (Z-score = 0.2). As the score cutoffs required by this comparison tool are E-value $\leq$ 100, Z-score > 0, and number of equivalent residue pairs $\geq$15, this particular query-library hit clearly falls just within the boundaries of the accepted score ranges.

The nuclease domain of putative ATP-dependent RNA helicase Hef from *Pyrococcus furiosus* (pdb|1j22, 1j23, 1j24, and 1j25 (Nishino, Komori et al. 2003); SCOP domains: d1j22a_, d1j23a_, d1j24a_, and d1j25a_), a member of the restriction endonuclease-like superfamily in SCOP, is incorrectly mapped to the FAD/NAD(P)-binding domain superfamily. This assignment is made because of a conservation pattern analysis hit to NADH-dependent ferredoxin reductase BphA4 from *Pseudomonas* strain KKS102 (pdb|1d7y (Senda, Yamada et al. 2000); SCOP domain: d1d7ya2). Although the core of both the query and the library representative is an α/β domain containing a 5-stranded β-sheet, the overall topology is not similar. This query-library pair hit is accepted because of the matrix-based conservation score of 0.32, which is based on the structural alignment of these two domains by DaliLite (Z-score = 3.7), while the score

cutoffs required by this comparison tool are matrix-based score $\geq$ 0.25 and DaliLite Z-score $\geq$ 2. Again, the scores for this hit fall near the boundaries of the accepted score ranges.

The proteolytically-cleaved peptide C from bovine lysosomal α-mannosidase (pdb|1o7d (Heikinheimo, Helland et al. 2003); SCOP domain: d1o7d.2) belongs to the galactose mutarotase-like superfamily in SCOP, but is incorrectly mapped to the "alpha-Amylases, C-terminal domain β-sheet domain" superfamily. This assignment is due to a hit identified by conservation pattern analysis to the C-terminal domain of neopullulanase from *Bacillus stearothermophilus* (pdb|1j0h (Hondoh, Kuriki et al. 2003); SCOP domain: d1j0ha2). Although the core of lysosomal α-mannosidase peptide C and the C-terminal domain of neopullulanase each form a β-sandwich-like fold, the topologies of these folds are different. The COMPASS-based conservation score for this query-library pair (0.52) is based on the structural alignment of the two domains by DaliLite (Z-score = 4.6). These scores fall just within the required ranges for acceptance by the conservation pattern comparison method (COMPASS-based conservation score $\geq$ 0.5 and DaliLite Z-score $\geq$ 2).

4.3.3 Comparison of tweaking and testing set results

Table 4.1 shows that the SCOPmap results are comparable for the tweaking set and the testing set. SCOPmap performance on the two test sets are nearly equivalent: 94.3% (tweaking set) *vs* 94.5% (testing set) correct assignments; 5.7% (tweaking set) *vs* 5.3% (testing set) false negative assignments; and 0.0% (tweaking set) *vs* 0.2% (testing set) false positive assignments. The most significant differences are in the results for the specific types of correct assignments: true positives with ranges accurate within 10 residues, true positives with ranges that are not accurate within 10 residues, and true negatives. These seemingly disparate results are predominantly reflections of inconsistencies in test set composition rather than in SCOPmap performance. More specifically, these variations are primarily due to the number of query domains that

belong to new SCOP superfamilies. The most obvious consequence is the fraction of each test set given true negative assignments (6.2% in tweaking set, 3.1% in testing set), which is directly dependent on the fraction of each test set that belongs to new SCOP superfamilies. If domains from new SCOP superfamilies are ignored, the apparent disparity in SCOPmap boundary definition accuracy is reduced. For example, if the entire test sets are considered, there is a 2.3% difference in the number of domains correctly assigned whose ranges are accurate within 10 residues of the SCOP-defined boundaries. However, when considering only domains that can potentially be mapped correctly (i.e. domains that belong to existing SCOP superfamilies), 86.8% of the tweaking set domains are correct assignments that are accurate within 10 residues, compared to 86.4% of the testing set domains. Similarly, 92.4% of all correctly assigned domains in the tweaking set are accurate within 10 residues, compared to 91.6% for the corresponding domains in the testing set.

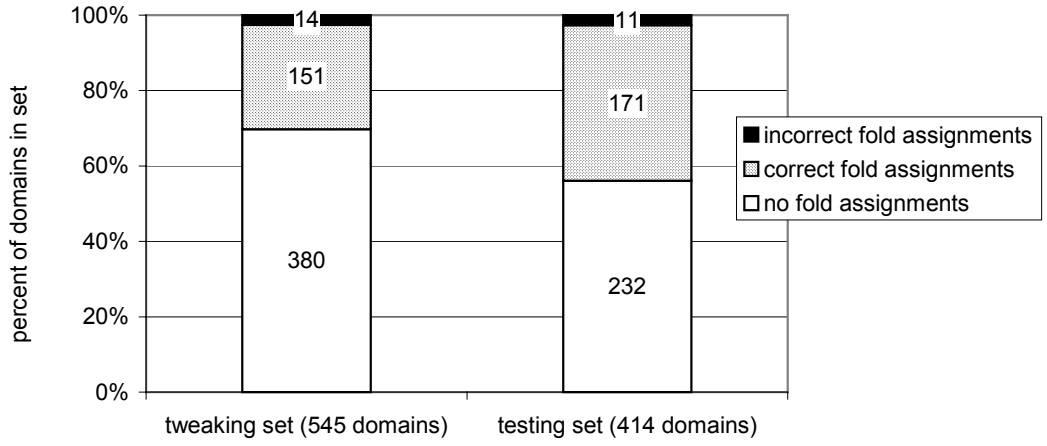
The comparable results are a reliable indication of the consistency of SCOPmap performance because the two test sets are of nearly equivalent difficulty. First, the two test sets include approximately the same fraction of trivial assignments: 73.7% of mappable domains in the tweaking set are assigned by gapped BLAST while 73.6% of mappable domains in the testing set are assigned by gapped BLAST. A "mappable domain" is defined as a domain that is both evolutionarily relevant and is a member of a SCOP superfamily that exists in the version of SCOP used as the library. Of the non-trivial mappable domains (i.e. mappable domains that are not assigned by gapped BLAST), the average sequence identity between the query domain and the closest library representative from the same SCOP superfamily is 29.2% in the tweaking set and 28.6% in the testing set.

4.3.4 Fold level assignments

Fold level assignments are attempted for regions of query chains at least 20 residues in length for which no superfamily assignment was made. Results are shown in

Figure 4.9. In the tweaking set (v1.61-v1.63 test set), fold level assignments are made for ~30% of the 545 SCOP-defined domains with no superfamily level assignment. 92% of these fold level assignments are correct. In the testing set (v1.63-v1.65 test set), fold level assignments are made for ~44% of the 414 SCOP-defined domains with no superfamily level assignment. Of these assignments, ~94% are correct.

Figure 4.9: Fold Level Assignments



Similar to the superfamily level assignments, the apparent disparity in fold level assignments are due primarily to the relative composition of the two test sets rather than inconsistency in performance. There are two principal attributes of test set composition that result in improved fold level results. First, domains from new folds are typically given no fold level assignment by SCOPmap, so a smaller fraction of unmapped domains from new folds will result in a decreased number of domains for which no assignment is made. Second, because the structural similarity between two domains from the same superfamily is likely to be greater than that between two domains from different superfamilies within the same fold, a larger fraction of unmapped domains from existing superfamilies will result in an increased number of correct fold level assignments. Both of these attributes favor the testing set over the tweaking set. This indicates that the testing set is less challenging in terms of fold level assignments, which is consistent with the improved results relative to the tweaking set (Figure 4.9).

Although no fold level assignment is made in a large number of cases (~70% of tweaking set unmapped domains and ~56% of testing set unmapped domains), this result is not altogether unexpected for several reasons. First, as discussed above, a significant fraction of the unmapped domains in each set belong to new SCOP folds, so no appropriate fold level assignment exists among the set of library representatives. Next, the minimum Z-score cutoff required for making fold level assignments is strict in order to minimize false positive assignments. Ortiz *et al.* report that MAMMOTH Z-scores greater than 5.25 are generally sufficient for fold predictions (Ortiz, Strauss et al. 2002); however, in reality, a MAMMOTH Z-score of 10 is required for making reliable fold assignments. Although 45% of domains in the tweaking set from existing folds but without a fold assignment (171 of 380 domains) have at least one MAMMOTH hit to a representative of the appropriate fold with a Z-score between 5.25 and 10, results in this range are not used due to many occurrences of false positive assignments. Conversely, because MAMMOTH Z-scores greater than 22 are sufficient for assignments at the superfamily level, fold assignments are neither necessary nor made for query-library domain pairs with such overwhelming structural similarity. Furthermore, because query-library domain pairs with sufficient sequence similarity to be recognized by automatic methods are mapped at the superfamily level, unmapped domains have very little sequence similarity to the corresponding library representatives. Consequently, fold assignments are made only for a rather limited set of queries: domains with extremely low sequence similarity as well as significant but not overwhelming structural similarity to library representatives.

The false positive rates are nearly identical in the two test sets (~2.6%). In both sets, the false positive rate of fold level assignments is significantly higher for domains that belong to new SCOP folds compared to those from existing SCOP folds. For example, in the second testing set, 6 of the 86 domains that belong to new folds have incorrect fold level assignments (7.0%) while only 5 of the 328 domains from existing folds are given an incorrect assignment (1.5%). Because false positive hits are likely to fall just above the Z-score cutoff for fold level assignment, many false positives are

ignored due to other hits found with better Z-scores, which are true positives in most cases. Thus, because domains that belong to existing SCOP folds should have significant structural similarity to at least one library domain (i.e. the library representative(s) of that particular SCOP fold), the negative effect of false positive hits to these domains is minimized in the false positive rate relative to that for domains from new SCOP folds.

False positive fold level assignments are typically due to a query and library representative sharing similar but not identical topology. For example, the structure of riboflavin kinase (pdb|1n06 (Bauer, Kemter et al. 2003); SCOP domain: d1n06b_) is a query in v1.61-v1.63 test set and belongs to a SCOP superfamily that is new to SCOP v1.63. Appropriately, no superfamily level assignment is made. The fold of riboflavin kinase is a n=6, S=10 β-barrel with strand order 163452, but SCOPmap assigns this domain to the double psi β-barrel fold in SCOP, which is an n=6, S=10 β-barrel with strand order 163425. In this case, the incorrect fold assignment is based on similarity of overall topology, but other false positive fold assignments occur when a region within a query domain and a region within a SCOP representative have similar topology despite overall dissimilarity of the folds. For example, the structure of the ε-subunit of the plasmid maintenance system (pdb|1gvn (Meinhart, Alonso et al. 2003); SCOP domain: d1gvna_) is another query in v1.61-v1.63 test set which also belongs to a new superfamily in SCOP v1.63. Again, no superfamily level assignment is made, as appropriate. The fold of the ε-subunit is a 3-helix up-and-down bundle with left-handed twist, but SCOPmap assigns this domain to a 4-helix up-and-down bundle fold. The three α-helices in the query domain and the last three α-helices of the SCOP representative have identical topology, similar lengths, and equivalent spatial orientation to each other. This false positive is a result of the query topology matching a region of a SCOP representative. The opposite case, when a region of the query domain is the same as the topology of an entire SCOP representative, occurs as well. For example, the structure of viral chemokine binding protein m3 (pdb|1mkf (Alexander, Nelson et al. 2002); SCOP domain: d1mkfa_), a query in v1.61-v1.63 test set, belongs to a new fold in SCOP v1.63. Appropriately, no superfamily level assignment is made for this query.

The fold of this domain is a 10-stranded β-sandwich with 6 β-strands in one β-sheet and 4 in the other. This domain is mapped at the fold level to an 8-stranded β-sandwich with 4 β-strands in each sheet. Although the overall folds of these two domains are different, 7 β-strands from these two β-sandwich folds have identical topology and mutual spatial arrangement.

Unsurprisingly, correct fold assignments are made predominantly for typical globular proteins while no fold assignments are made for small protein or coiled coil folds. Outside of this observation, there are no recognizable trends suggesting types of folds for which assignments are more easily made.

Furthermore, it should be noted that fold assignments are not the main goal of this algorithm. Rather, these assignments are a by-product of the comparison tools that are used for mapping at the superfamily level. The purpose of making fold level assignments is merely to assist the user in further study of those domains for which SCOPmap does not give a superfamily level assignment. The fold level mapping strategy and score cutoffs have not been optimized for high sensitivity or low false positives.

4.3.5 Performance of SCOPmap compared to SUPERFAMILY

SUPERFAMILY is another tool that attempts to assign domains within a query protein to the superfamily level of SCOP. The results of the performance of SUPERFAMILY relative to SCOPmap are shown in Table 4.1. Overall, SCOPmap performs better than SUPERFAMILY. SUPERFAMILY correctly maps 91.4% of domains compared to the 94.3% assigned to the correct SCOP superfamily by SCOPmap. Furthermore, SCOPmap is much more proficient at defining accurate domain boundaries. SCOPmap delineates domain boundaries within 10 residues of the SCOP-defined boundaries for 81.4% of domains, while SUPERFAMILY performs as well in only 70.1% of cases. This difference is due partly to the use of structural comparison tools such as MAMMOTH and DaliLite in the SCOPmap algorithm. However, the results of this algorithm when using only sequence comparison tools show that there is still a 6.5%

advantage over SUPERFAMILY in terms of accurately defined ranges (Table 4.1). Thus, the inclusion of structure comparison methods is not solely responsible for the dramatic improvement in boundary definition. Presumably, a second predominant factor in the increased domain boundary accuracy is the strict coverage criteria for sequence comparison methods incorporated in SCOPmap.

Table 4.1 shows the results of using only the BLAST, RPS-BLAST, PSI-BLAST, and COMPASS portions of this algorithm. This modified version of SCOPmap (henceforth referred to as the "sequence-only algorithm") was expected to perform similarly, if not better than, SUPERFAMILY. It was therefore surprising to observe significantly more false negative assignments by the sequence-only algorithm compared to the SUPERFAMILY algorithm (12.5% and 8.6%, respectively). Investigation of the 573 false negatives from the sequence-only algorithm indicates three general explanations for these missed assignments. In ~47% of these cases (270 of 573 domains), there are no sequence comparison hits below the required E-value thresholds. Next, in ~17% of cases (97 of 573 domains), sequence hits that pass both the E-value and coverage criteria are found, but the domain is not assigned due to an unresolved choice between conflicting superfamilies. In the remaining 36% of cases (206 of 573 domains), sequence comparison hits to at least one superfamily representative are found that pass the required E-value cutoffs but fail the coverage criteria. These 206 domains correspond to ~4.5% of this test set and account for the difference in false negative rates between the sequence-only algorithm and SUPERFAMILY, which does not have a coverage requirement.

4.3.6 SCOPmap and SUPERFAMILY performance on non-trivial domain assignments

Because nearly 70% of the domains can be mapped using gapped BLAST (Table 4.3), the results of both SCOPmap and SUPERFAMILY are skewed in favor of trivial domain assignments. In order to evaluate the performance of these two programs on

more challenging assignments, the results were re-tabulated excluding all domains assigned via gapped BLAST (Table 4.2). Here, SCOPmap assigns 81.6% of domains to the appropriate SCOP superfamily while SUPERFAMILY correctly maps 77.1% of domains, so SCOPmap's advantage in correctly assigned domains increases from 2.9% for all domains to 4.5% for only non-trivial assignments. SCOPmap's proficiency in domain boundary definition is also accentuated, as the difference in percent of domains with accurately defined domain boundaries increases from 11.3% for all domains (SCOPmap: 81.4%, SUPERFAMILY: 70.1%) to 12.8% for non-trivial assignments (SCOPmap: 42.8%, SUPERFAMILY: 30.0%). Thus, evaluating only the non-trivial assignments emphasizes the advantages of SCOPmap over SUPERFAMILY.

Table 4.2: Automatic Mapping Results for Non-trivial Assignments

Result	*SCOPmap*		*SUPERFAMILY*	
	# of domains	*% of test set*	*# of domains*	*% of test set*
Assignment to correct SCOP superfamily, boundaries within 10 residues	607	**42.8%**	425	**30.0%**
Assignment to correct SCOP superfamily, boundaries not within 10 residues	252	**17.8%**	379	**26.7%**
Domain belongs to a new SCOP superfamily, no assignment made	284	**20.0%**	241	**17.0%**
Acceptable assignment, but not the same assignment as given in SCOP	13	**0.9%**	48	**3.4%**
Domain belongs to an existing SCOP superfamily, no assignment made	261	18.4%	324	22.9%

Table 4.2: Automatic Mapping Results for Non-trivial Assignments. "Non-trivial domains" are the 1417 domains could not be assigned by gapped BLAST. Boldface text indicates correct assignments.

4.3.7 False negative assignments by SCOPmap and SUPERFAMILY

The false negative assignments made by SCOPmap (261 domains) and by SUPERFAMILY (395 domains) were compared in order to determine the degree of overlap between the two sets of unassigned domains. One might expect that a significant

number of the false negative assignments would be shared by the two algorithms and would represent those cases that are too difficult to be confidently mapped by existing automatic comparison tools. Indeed, 205 domains are given false negative assignments by both SCOPmap and SUPERFAMILY.

Therefore, of the 261 false negative assignments made by SCOPmap, only 56 domains (21%) are correctly mapped by SUPERFAMILY. 38 of these domains were correctly identified by at least one of the comparison methods used but were not assigned (due, for example, to an unresolved choice of superfamily assignment). Most of the remaining domains that were assigned by SUPERFAMILY but not identified by SCOPmap represent cases that are typically difficult for automatic methods: 8 are small disulfide-rich domains, 3 are relatively short domains (74, 75, and 126 residues) that are interrupted by very large insertions (290, 289, and 282 residues respectively), and 1 domain contains many short breaks in the sequence and structure. The few remaining examples are domains that could have been reasonably expected to be mapped by SCOPmap: *E. coli* succinate dehydrogenase subunit SdhC (pdb|1nek chain D (Yankovskaya, Horsefield et al. 2003) and pdb|1nen chain D (Yankovskaya, Horsefield et al. 2003); SCOP domains: d1nekd_ and d1nend_) is a helical bundle protein that belongs to the succinate dehydrogenase/fumarate reductase transmembrane segment superfamily in SCOP, and the PKD-like domain of *Methanosarcina mazei* surface layer protein (pdb|1l0q (Jing, Takagi et al. 2002), chains A, B, C, and D; SCOP domains: d1l0qa1, d1l0qb1, d1l0qc1, d1l0qd1) is an immunoglobulin-like domain that belongs to the PKD domain superfamily in SCOP. Other than the low sequence identity between these queries and the library representatives of the corresponding SCOP superfamilies, there are no convincing arguments for why these assignments might not be made. In each of these cases, significant hits are found by the structure comparison tools used in SCOPmap: SdhC has a DaliLite Z-score of 8.7 to a library representative of its SCOP superfamily, and surface layer protein PKD-like domain has a MAMMOTH Z-score of 10.6 to the library representative of its SCOP superfamily. However, the limited sequence similarity between the query and representative domains results in insufficient

BLOSUM scores to meet the required score cutoffs of these methods. Although these are consequently false negative assignments at the superfamily level, the correct fold level assignment was made in each of these last 6 cases.

Conversely, approximately half of the false negative assignments made by SUPERFAMILY (190 of 395 domains) are correctly mapped by SCOPmap. Of these domains, ~54% are first identified by a sequence comparison tool in SCOPmap (gapped BLAST, RPS-BLAST, PSI-BLAST, or COMPASS), ~29% are first identified by a structure comparison tool (MAMMOTH or DaliLite), and the remaining ~17% are first identified by a method that combines both sequence and structure information (correlation of conservation patterns or the agreement of DaliLite alignments with gapped BLAST, RPS-BLAST, or PSI-BLAST alignments).

4.4 DISCUSSION

4.4.1 Performance of individual comparison methods

In order to assess the relative performance of the individual comparison tools used by SCOPmap, the number of assignments in the tweaking set gained by each additional comparison method was evaluated. The results are summarized in Table 4.3. For each comparison tool, the number of domains first identified by that method was determined, and the percent of previously unassigned domains gained by that method was calculated. The comparison tools are listed in order of increasing sensitivity to distant homologs: sequence comparison methods (BLAST, RPS-BLAST, PSI-BLAST, and COMPASS), structure comparison methods (MAMMOTH and DaliLite), and finally comparison methods that incorporate both sequence and structure information (correlation of conservation patterns and agreement of DaliLite alignments with BLAST, RPS-BLAST, or PSI-BLAST alignments). Domains are included in the total count for only the least sensitive comparison tool that identified the hit.

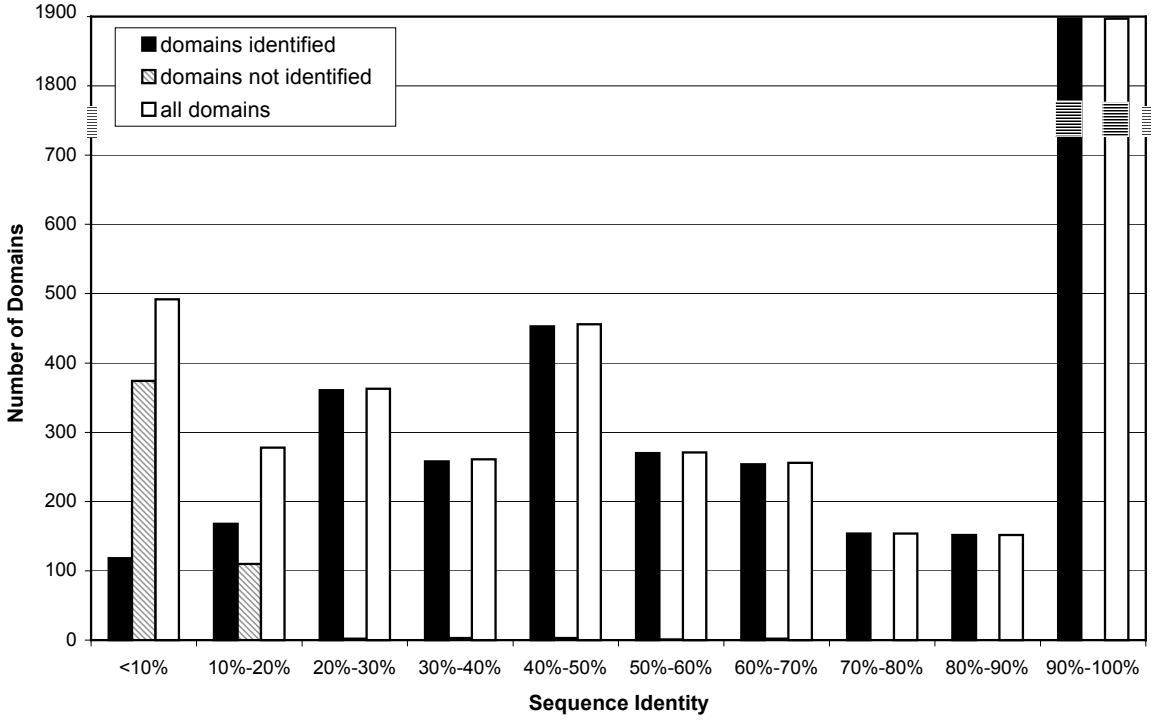
Table 4.3: Domain Assignments by Increasingly Sensitive Comparison Methods

Comparison Method	*Number of Domains First Identified By This Method* [4035 mapped domains plus 50 domains that are identified but not assigned (see Table 4.4)]	*Average Sequence Identity Between Query and Closest Superfamily Representative*	*% of Domains Unmapped by Less Sensitive Methods that are Identified by This Method*
BLAST	3163	80.1%	69.1%
RPS-BLAST	514	41.1%	36.3%
PSI-BLAST	104	26.1%	11.5%
COMPASS	26	27.2%	3.3%
MAMMOTH	100	29.7%	12.9%
DaliLite	124	17.4%	18.4%
correlation of conservation patterns	23	11.1%	4.2%
agreement of alignments produced by DaliLite and by gapped BLAST, RPS-BLAST, or PSI-BLAST	31	12.1%	5.9%

The greatest number of assignments are made by gapped BLAST and RPS-BLAST, which give 69.1% gain and 36.3% gain of previously unmapped domains, respectively. However, these assignments are among the easiest in the set. The average sequence identity between the query domain and the closest library representative of that superfamily is 80.1% for gapped BLAST assignments and 41.1% for RPS-BLAST assignments. Furthermore, these numbers are considerably inflated as a consequence of the surfeit of trivial assignments in the tweaking set (Figure 4.10).

PSI-BLAST, MAMMOTH, and DaliLite each give between 10% and 20% gain of previously unmapped domains. The average sequence identities between the identified query domains and the library domains indicate that these assignments are neither trivial nor unusually difficult. The two structure comparison methods show similar overall performance by this assessment, although DaliLite does have the advantage over MAMMOTH both in number of assignments and percent gain as well as in difficulty of assignments made. This seemingly implies that comparison via MAMMOTH is an

Figure 4.10: Sequence Identity Between Tweaking Set Domains and the Closest Library Representative From the Same SCOP Superfamily



unnecessary step, and indeed, nearly all domain assignments made by MAMMOTH are also made by DaliLite. However, MAMMOTH is a much faster than DaliLite. Furthermore, MAMMOTH is both necessary for and proficient at determining potential hits by DaliLite. The pre-identification of potential hits drastically reduces the running time compared to comprehensive comparison of the query domains to all library domains by DaliLite. Furthermore, MAMMOTH is essential for making fold level assignments.

The conservation pattern analysis and the calculation of agreement between DaliLite alignments and BLAST, RPS-BLAST, or PSI-BLAST alignments have 4.2% and 5.9% gain of previously unmapped domains, respectively. Although the numbers of additional assignments are among the lowest of any of the comparison tools, these two methods also make the most challenging assignments of any of the comparison tools included in SCOPmap. The average sequence identity between query domains and

library representatives for assignments made first by these methods is less than 15%. Specific examples are discussed in section 4.4.2.

Thus, the general observation is that, as expected, those comparison tools more sensitive to distant homology typically make more challenging assignments, but with lower percent gains. The only clear exception to this trend is COMPASS. COMPASS has the lowest percent gain of any step at 3.3%, and the domains first identified by this method are only moderately difficult assignments (average sequence identity 27.2%). This is presumably due in part to the extremely strict E-value cutoff necessary for avoiding false positives ($1x10^{-10}$). Furthermore, of the four sequence comparison tools used in SCOPmap, COMPASS is most sensitive to remote homologs. Therefore, if the query-library domain pair has sufficient sequence similarity to be recognized by automatic methods, it is likely that the hit would also be identified by a less sensitive sequence comparison tools and consequently be accounted for earlier in Table 4.3.

4.4.2 SCOPmap performance on remote homologs

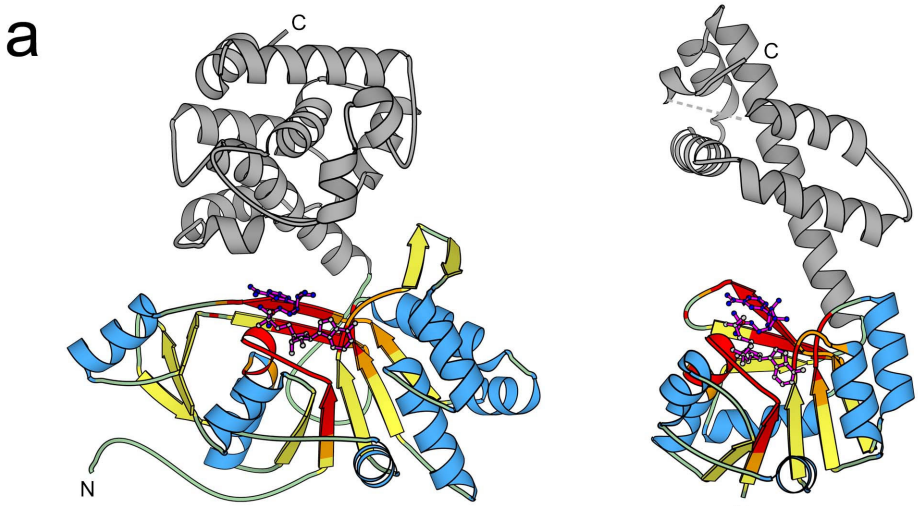
Correctly mapped remote homologs

The similarity of the tweaking set to the representative library domains is shown in Figure 4.10 (white bars). Nearly 50% of tweaking set domains are more than 70% identical to one of the library representatives from the same SCOP superfamily. Furthermore, 69.1% of the tweaking set domains can be correctly mapped by gapped BLAST (Table 4.3). Other domains, however, are more difficult to assign due to limited similarity of the query domain to the representative library domains. SCOPmap is able to make several such assignments, including nearly 300 domains with less than 20% sequence identity to the closest library domain from the same SCOP superfamily (black bars, Figure 4.10).

A prevalent difficulty in classifying proteins with automatic methods is correctly assigning domains with only limited sequence similarity to library representatives. One such example of a difficult but correctly assigned domain is the N-terminal domain of

mannitol 2-dehydrogenase from *Pseudomonas fluorescens* (pdb|1lj8 (Ortiz, Strauss et al. 2002); SCOP domain: d1lj8a2). In SCOP, this domain belongs to the NAD(P)-binding Rossmann-fold domains superfamily. There are 90 representatives of this superfamily in the library, all of which have less than 10% sequence identity to the query domain. There are no BLAST, RPS-BLAST, PSI-BLAST, COMPASS, MAMMOTH, or DaliLite hits to these library representatives that pass both the required coverage and E-value or Z-score thresholds. Hits to three of the 90 superfamily representatives are identified by DaliLite: the N-terminal domain of glycerol-3-phosphate dehydrogenase from *Leishmania mexicana* (pdb|1evy (Suresh, Turley et al. 2000); SCOP domain: d1evya2) with Z-score 6.9, the N-terminal domain of conserved hypothetical protein MTH1747 from *Methanobacterium thermoautotrophicum* (pdb|1i36 (Korolev, Dementieva et al. *To be published*)) with Z-score 6.3, and the N-terminal domain of lactate/malate dehydrogenase from *Methanococcus jannaschii* (pdb|1hye (Lee, Chang et al. 2001); SCOP domain: d1hyea1) with Z-score 6.4. Because of the poor BLOSUM scores calculated for the pairwise alignments given by DaliLite, none of these hits are accepted by the DaliLite comparison method. However, these relatively high Z-scores indicate that the DaliLite alignments are reliable enough for use in the comparison of conservation patterns method, and hits to two of these superfamily representatives are accepted based on correlation of conservation patterns: the N-terminal domain of glycerol-3-phosphate dehydrogenase (SCOP domain: d1evya2) has matrix-based conservation score = 0.26, and the N-terminal domain of conserved hypothetical protein MTH1747 (SCOP domain: d1i36a2) has matrix-based conservation score = 0.11. In both of these cases, approximately 75% of the most conserved positions in the query domain and in the library domain are equivalent (Figure 4.11b). Furthermore, these most conserved positions are clustered around the nucleotide-binding sites, which are equivalent in these domains (Figure 4.11a). The N-terminal domain of this query structure is therefore mapped to the NAD(P)-binding Rossmann-fold domain superfamily in SCOP based on the high degree of correlation between the conservation patterns of the query domain and these two superfamily representatives.

Figure 4.11: Correctly Mapped Remote Homolog: N-terminal Domain of Mannitol 2-dehydrogenase



b

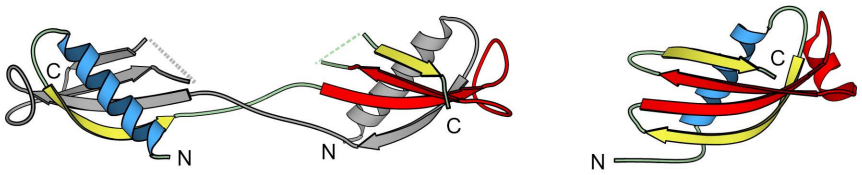
```
              EEEEE   HHHHHHHHHHHHHHHHHH    EEEEEEEE   HHHHHHHHHHHH
1lj8A   24 dtrQGIAHIGVGGFHRAhQAYYTDAlmntgegLDWSICGVGLRseDRKARddlagqd[18]IG
1i36A    1 ---LRVGFIGFGEVAQT-LASRLRS-------RGVEVVTSLEG--RSPSTierar------TV
             **********

             EEEE   HHHHHHH   EEEEE              HHHHHHH     EEEEEEE
1lj8A  101 SISDM[9]LIDKLASpeIRIVSLTITEGGY[25]pktVFGFICA[10]pAFTVMSCDNLPHn-
1i36A   45 GVTET---SEEDVYS--CPVVISAVTPGVA----lgaARRAGRH----vRGIYVDINNISPet
                             **                            ****

           HHHHHHHHH    EEEEEEEEEE            EEEEEE
1lj8A  196 GAVTRKALL[16]HVSFPNAMVDRIT[24]CEPF--VQWVLEDKFvngrpAWEK--VGVQFTD 286
1i36A   95 VRMASSLIE----KGGFVDAAIMGSV----RRKGadIRIIASGRDaeefmKLNRygLNIEVRG 149
                        *****        *     ***
```

Figure 4.11: Correctly Mapped Remote Homolog: N-terminal Domain of Mannitol 2-dehydrogenase. a) MOLSCRIPT diagrams of mannitol 2-dehydrogenase from *Pseudomonas fluorescens* (left, pdb|1lj8) and library representative conserved hypothetical protein MTH1747 from *Methanobacterium thermoautotrophicum* (right, pdb|1i36). The N-terminal domains, which belong to the NAD(P)-binding Rossmann-fold superfamily, are shown in color. Regions in red are positions among the top 25% of most conserved positions in both the query (1lj8A, N-terminal domain) and library representative (1i36A, N-terminal domain). Regions in orange are positions among the top 25% of most conserved positions in either the query or the library representative domain, but not both. Positions in this domain that are not among the most highly conserved are blue (α-helices), yellow (β-strands), and green (coils). The C-terminal domain is shown in grey, dashed lines indicate disordered regions, and the bound nucleotide is shown in ball-and-stick format and is colored magenta. b) Pairwise alignment of the query (1lj8A) and library representative (1i36A) from DaliLite results. Residues in red bold text are among the top 25% of most conserved positions in at least one of the domains. Residues indicated with an asterisk are among the top 25% of most conserved positions in both the query and library domains. Secondary structure is indicated above the alignment, with E signifying β-strand residues and H signifying α-helix residues. The numbers flanking the alignment indicate the residue number in the sequence of the first (or last) aligned residue on that line. Numbers in brackets specify the number of residues in an insertion that are not shown. Capital letters are residues aligned by DaliLite and lower-case letters are unaligned residues.

Conformational differences between similar protein domains also result in challenging classification assignments for automatic structure comparison tools. One such example is the antimicrobial cathelicidin motif of protegrin-3 from *Sus scofa* (pdb|1lxe (Sanchez, Hoh et al. 2002); SCOP domain: d1lxea_). The crystal structure of this protein shows the domain in a swapped dimer conformation (Figure 4.12a, left). The closest library representative to this query domain is cystatin from *Gallus gallus* (pdb|1cew (Bode, Engh et al. 1988); SCOP domain: d1cewi_), which belongs to the cystatin/monellin superfamily in SCOP. This domain is a monomer in the crystal structure (Figure 4.12a, right). The sequence identity between the query (cathelicidin motif of protegrin-3) and this library representative (cystatin) is approximately 19%. The

Figure 4.12: Correctly Mapped Domain with Conformational Variation: Cathelicidin Motif of Protegrin-3

a



b

RPS-BLAST alignment (E-value: 16)

```
1lxeA    5 YREAVLRAVDRLNEQSSEANLYRLLELDQPPKADEDPGTPKPVSFTVKETVCPRPTRQPP--
             + A+  A+   N  S++    R++ +    +          +   +  T CP+ +
1cewI   13 LQRALQFAMAEYNRASNDKYSSRVVRVISAKRQLVSGIKYI-LQVEIGRTTCPKSSGDLQSC

1lxeA   65 ELCDFKENGRVKQCVGTV 82
           E  D  E  +   C   V
1cewI   74 EFHDEPEMAKYTTCTFVV 91
```

DaliLite alignment (Z-score: 2.4)

```
1lxeA   40 DPGTPKPVSFTVKETVCPRPTR--QPPELCDfkENGRVKQCVGTVTLD---qikdplDITCN 96
1cewI   47 VSGIKYILQVEIGRTTCPKSSGdlQSCEFHDepEMAKYTTCTFVVYSIpwlnqikllESKCQ 108
```

Figure 4.12: Correctly Mapped Domain with Conformational Variation: Cathelicidin Motif of Protegrin-3. a) The cathelicidin motif of protegrin-3 from *Sus scofa* (left, pdb|1lxe) is in a swapped dimer conformation. One monomer in the complex is colored, and the second monomer is grey. Disordered regions are indicated by dashed lines. Cystatin from *Gallus gallus* (right, pdb|1cew) is a library representative of the cystatin/monellin superfamily. b) Pairwise alignments of this query (1lxeA) and library (1cewI) domain produced by RPS-BLAST and DaliLite. Residues aligned equivalently by these two comparison tools are in red bold. The equivalently aligned regions are shown in red in the structure figures. In the DaliLite alignment, capital letters are aligned residues and lower-case letters are unaligned residues.

hit between the query and this library representative is found by both the RPS-BLAST and DaliLite methods. However, the scores for these hits are relatively poor as a result of the low sequence identity and the conformational variation between the two domains. The scores for these comparisons (RPS-BLAST E-value = 16 and DaliLite Z-score = 2.4) fail the score cutoff criteria for these methods individually. Comparison of the alignments produced by these two methods, however, indicates that a significant portion of the domain is aligned equivalently by RPS-BLAST and DaliLite (Figure 4.12b). Thus, based on the agreement of these two methods, the cathelicidin motif of protegrin-3 is correctly mapped to the cystatin/monellin superfamily of SCOP.

Another common problem for many automatic comparison methods is the presence of large insertions (or deletions) in the query domain. This third example demonstrates the ability of the mapping program to correctly assign such cases. Monomeric isocitrate dehydrogenase from *Azotobacter vinelandii* (pdb|1itw (Yasutake, Watanabe et al. 2002); SCOP domain: d1itwa_) belongs to the isocitrate/isopropylmalate dehydrogenase superfamily in SCOP. There are two representatives of this superfamily in the library, both of which have less than 15% sequence identity to the query domain. Furthermore, the query domain has an approximately 250-residue insertion relative to the superfamily representatives (Figure 4.13). There are no BLAST, RPS-BLAST, PSI-BLAST, or COMPASS hits to either library representative. Although the MAMMOTH hit to 3-isopropylmalate dehydrogenase from *Salmonella typhimurium* (pdb|1cnz (Wallon, Kryger et al. 1997); SCOP domain: d1cnza_) is accepted with Z-score 22.2, the presence of the large insertion in the query results in an erroneous range definition by MAMMOTH (Figure 4.13b). Comparison of the query to this same library representative by DaliLite identifies residues 164-397 as an insertion in this domain (Figure 4.13b). Although SCOP assigns the entire chain of monomeric isocitrate dehydrogenase as one domain (residues 1-741), residues 150-404 are defined as an insert region. Thus, the DaliLite-based assignment made by SCOPmap (residues 2-163, 398-671) is a reasonably accurate domain definition.

Figure 4.13: Correctly Mapped Domain with Large Insertion: Monomeric Isocitrate Dehydrogenase

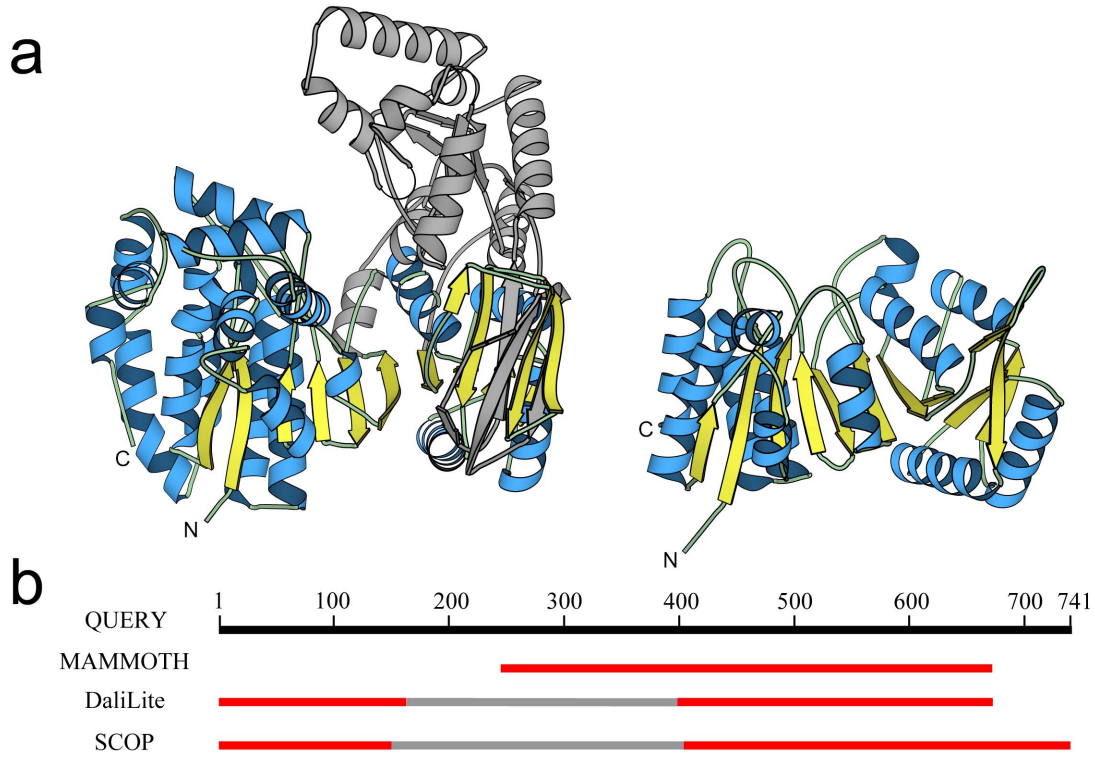


Figure 4.13: Correctly Mapped Domain with Large Insertion: Monomeric Isocitrate Dehydrogenase. a) MOLSCRIPT diagram of monomeric isocitrate dehydrogenase from *Azotobacter vinelandii* (left, pdb|1itw). The insert region as defined by SCOP is shown in grey. Isopropylmalate dehydrogenase from *Salmonella typhimurium* (right, pdb|1cnz) is a library representative of the isocitrate/isopropylmalate dehydrogenase superfamily. b) Range assignments as made by MAMMOTH, DaliLite, and SCOP. Regions assigned to the isocitrate/isopropylmalate dehydrogenase superfamily are red, insert regions are grey.

Domains without SCOPmap assignments at the superfamily level

In 5.7% of the tweaking set, no superfamily assignment is made for domains that belong to superfamilies included in SCOP v1.61. General explanations for these false negative assignments are summarized in Table 4.4. Of the 261 unmapped domains, 19.2% (50 domains) are found by meeting the required score cutoffs of one or more of the comparison tools used, but these domains are not assigned due to a conflict with

another domain identified in the same query chain. There are two ways in which this may happen: there may be an unresolved choice of superfamily assignment over a certain region of the query chain, or the boundary of one domain may erroneously extend over a second domain resulting in the assignment of one domain while the other is missed.

In the remaining 80.8% of unmapped domains, comparison of the query to the library domains do not pass the score cutoffs of any of the methods used. These domains typically have only limited structural similarity as well as less than 20% sequence identity to the library representatives. All domains that have greater than ~20% sequence identity to a library representative from the same SCOP superfamily but are not identified by any of the comparison tools used in SCOPmap are small protein domains less than 50 residues in length. Because automatic methods often perform poorly on small proteins, such cases are not unexpected. These unmapped small protein examples comprise only 0.2% of the tweaking set. Furthermore, the unmapped domains often have inserted or deleted structural elements relative to the library domains. The unmapped and unidentified domains fall into three general categories in terms of structural similarity to the library representatives. First, 33.3% of unmapped domains have very little structural similarity to the corresponding library domains. When the MAMMOTH scores for a query domain are insufficient for making superfamily assignments, these scores are used

Table 4.4: SCOP Domains Unassigned by SCOPmap at the Superfamily Level

Whether Domain is Identified by at Least One Comparison Method	*Reason Domain is Unmapped*	*Number of Domains*	*% of Unassigned Domains*	
The domain is identified by one or more methods, but is not assigned.	The boundary assigned to one domain in the query chain is extended too far and, as a result, a second domain assignment is missed.	22	8.5%	19.2%
	Unresolved choice between conflicting superfamilies.	28	10.7%	
Domain is not identified by any comparison tool used in SCOPmap.	DaliLite hits to superfamily representatives fail "accepted hit" cutoffs.	108	41.4%	80.8%
	At least one superfamily representative identified as potential hit via MAMMOTH, but DaliLite produces no output for the comparison.	16	6.1%	
	No superfamily representatives have MAMMOTH scores high enough to be identified as potential hits via DaliLite.	87	33.3%	

as an initial indicator of whether specific query-library domain pairs are likely to be assigned by DaliLite (see section 4.2.2 for a detailed explanation). For unmapped domains, the MAMMOTH scores to library domains are too poor to be identified even as potential hits. Next, there are a small number of cases (6.1% of unmapped domains) that have potential but unconfirmed structural similarity to library representatives. In these cases, one or more potential hits are identified by MAMMOTH, but DaliLite does not produce output for those pairs. This could mean that the DaliLite Z-score is less than zero for the given pair of domains, or that either the query domain, the library representative, or both could not be handled by DaliLite because, for example, the structure lacks recognizable secondary structure, contains only C_α coordinates, is less than 30 residues in length, *etc*. Finally, the remaining 41.4% of unmapped domains have recognizable but insufficient structural similarity to the library representatives. For these domains, hits are found via DaliLite but the scores of the hits do not meet the required cutoffs. Because such scores cannot be confidently distinguished from false positives, no superfamily assignment is made.

Since the inception of the SCOP database, the rapid growth in the number of available protein structures has resulted in a classification scheme that is not equally uniform in all parts. This is primarily apparent in overpopulated folds and superfamilies, such as TIM β/α-barrels, where intermediate relationships exist but are difficult to describe within the original SCOP classification scheme. These special cases in the SCOP database also contribute to the rate of false negative assignments by SCOPmap. In a later section, the conservative nature of SCOP is demonstrated by cases in which homologous proteins are assigned to different superfamilies. As a consequence of this attribute of the SCOP database, good hits via automatic comparison methods are sometimes found to multiple SCOP superfamilies. In some cases, SCOPmap is not capable of selecting one final assignment out of several correct choices. These 28 examples, which make up the unresolved choice of superfamilies category in Table 4.4, account for less than 1% of the tweaking set but 10.7% of all false negative assignments. Conversely, there are also numerous instances in which the SCOP classification is quite

liberal. Examples are rampant in the sections of the database that the authors describe as not a part of the proper SCOP classification, such as the low resolution structures and peptides classes. These classes are not included in the SCOPmap library and are therefore not considered by the SCOPmap algorithm. However, cases were also observed in the evolutionarily relevant multi-domain proteins class of SCOP. The multi-domain proteins class is problematic in the sense that it deviates from the format followed by the remainder of the SCOP database. Members of this class have not been classified at the domain level, and there is often wide variation in the size and domain composition of the entries. One such example was detected during the manual investigation of false negative assignments from the tweaking set. Reovirus polymerase λ3 (pdb|1n1h (Tao, Farsetta et al. 2002); SCOP domain: d1n1ha_) belongs to the DNA/RNA polymerases superfamily in the multi-domain proteins class of SCOP. The structural fold of domains in the DNA/RNA polymerases superfamily has been described as a "right-hand" configuration containing "palm", "fingers", and "thumb" subdomains. Domains in this superfamily, of which there are >200, typically include 2 or 3 subdomains of the "right-hand" fold. For example, Moloney murine leukemia virus (MMLV) reverse transcriptase (pdb|1mml (Georgiadis, Jessen et al. 1995); SCOP domain: d1mml__), which is one of the representatives of this superfamily included in the v1.61 library, is a 265-residue fragment containing only the "palm" and "fingers" subdomains. Reovirus polymerase λ3, however, also includes a 380-residue N-terminal domain as well as a 377-residue C-terminal "bracelet" domain, in addition to the "palm", "fingers", and "thumb" subdomains. Thus, a 1267-residue, 3-domain protein (reovirus polymerase λ3) and a 265-residue, single domain fragment (MMLV reverse transcriptase) are classified equivalently at the superfamily level in SCOP. Naturally, such variations within the database are problematic for making appropriate classifications via automatic methods.

Examples of false negative SCOPmap assignments

Some superfamily assignments are missed due to extremely limited similarity between the query domain and the corresponding library representatives. One such

example is *Saccharomyces cerevisiae* DNA-binding domain from transcription factor Ndt80 (pdb|1mnn (Lamoureux, Stuart et al. 2002); SCOP domain: d1mnna_), which belongs to the p53-like transcription factors superfamily in SCOP. Members of this superfamily bind DNA through an s-type Ig fold. There are seven library representatives of this superfamily, all of which have less than 10% sequence identity with the query domain. There are no hits to these representatives found by BLAST, RPS-BLAST, or PSI-BLAST with E-value less than 100 or by COMPASS with E-value less than $1x10^{-3}$. Because the MAMMOTH hits to these representatives are very poor (Z-scores below 2.5), MAMMOTH finds neither accepted hits nor potential hits for comparison via DaliLite. Although the conserved core of this superfamily is observable by eye (Figure 4.14a), the many inserted structural elements relative to the library representatives contribute to the poor performance of the automatic structural comparison methods. The DNA-binding function of this domain may have contributed to its inclusion in this superfamily by the SCOP authors.

Superfamily assignments are also missed in cases where the similarity to library representatives is moderately significant but still insufficient for distinction from false positives. One such example is adaptor protein ClpS from *E. coli* (pdb|1lzw (Zeth, Ravelli et al. 2002), chain A; SCOP domain: d1lzwa_) (Figure 4.14b), which belongs to the ClpS-like superfamily in SCOP. The one representative of this superfamily in the library shares ~11% sequence identity with the query domain. BLAST, RPS-BLAST, and PSI-BLAST hits to this library representative are not found with E-values below 100, and a COMPASS hit to the library domain is not found with E-value below $1x10^{-3}$. Comparison of the query and library domain by MAMMOTH and DaliLite give more substantial results: a Z-score (M_Z) of 10.4 with BLOSUM score $-1.0x10^{-2}$ for the pairwise alignment produced by MAMMOTH and a Z-score (D_Z) of 8.8 with BLOSUM score $4.5x10^{-4}$ for the pairwise alignment produced by DaliLite. Unfortunately, these scores fall just below the required cutoffs for superfamily assignment via these methods. Thus, no superfamily assignment is made. However, the MAMMOTH Z-score does meet the fold level cutoff, so a correct fold assignment is made for this query domain.

Figure 4.14: Examples of False Negative SCOPmap Assignments

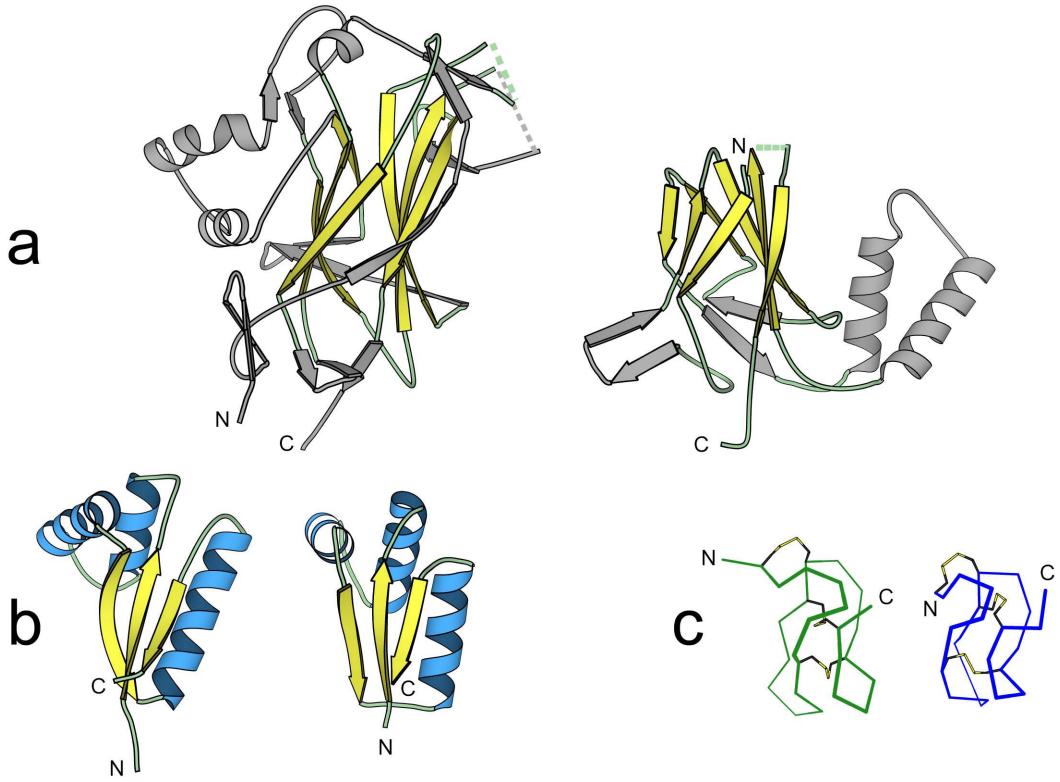


Figure 4.14: Examples of False Negative SCOPmap Assignments. a) MOLSCRIPT diagrams of unmapped domain transcription factor Ndt80 (left, pdb|1mnn) and library representative p52 subunit of NF-kappa B, N-terminal domain (right, pdb|1a3q, residues A37-A226). β-strands that belong to the Ig fold core are yellow, and additional structural elements are grey. Dashed lines indicate disordered regions. b) MOLSCRIPT diagrams of unmapped domain *E. coli* adaptor protein ClpS (left, pdb|1lzw, chain A) and library representative ribosomal protein L7/12 from *E. coli*, C-terminal domain (right, pdb|1ctf). c) C_α traces of unmapped domain δ-conotoxin TxVIA from *Conus textile* (left, pdb|1fu3) and library representative ω-conotoxin TxVII from *Conus textile* (right, pdb|1f3k). These two conotoxins share ~40% sequence identity. Disulfide bonds are shown in ball-and-stick format.

Additionally, technical shortcomings of automatic methods contribute to missed superfamily assignments. For example, δ-conotoxin TxVIA from *Conus textile* (pdb|1fu3 (Kohno, Sasaki et al. 2002); SCOP domain: d1fu3a_) is a 27-residue small protein that belongs to the omega toxin-like superfamily in SCOP. There are 21 library representatives of this superfamily, some of which share up to 40% sequence identity with the query domain. However, there are no hits to these representatives found by BLAST, RPS-BLAST, or PSI-BLAST with E-value less than 100 or by COMPASS with

E-value less than $1x10^{-5}$. The MAMMOTH hits to these 21 representatives all have Z-scores well below 4. Furthermore, DaliLite cannot analyze this protein due to the short length, thus precluding DaliLite comparisons with library representatives. Thus, despite significant sequence and structural similarity of δ-conotoxin TxVIA to several library representatives (Figure 4.14c), no superfamily assignment is made due to the poor performance of automatic methods on small proteins.

4.4.3 Finding new links between SCOP superfamilies: examples of homologs in different SCOP superfamilies identified by SCOPmap

The thiamin phosphate synthase superfamily and the ribulose-phosphate binding barrel superfamily are one example of homologous SCOP superfamilies identified by SCOPmap. Both superfamilies have a TIM β/α-barrel fold. When thiamin phosphate synthase is used as the query, hits to 8 different members of the ribulose-phosphate binding barrel superfamily are identified. These hits are found by PSI-BLAST, COMPASS, DaliLite, and the agreement between pairwise alignments produced by DaliLite and by RPS-BLAST or PSI-BLAST. Because confident hits are identified by both sequence and structure comparison methods, the homology between the two superfamilies is considered reliable, despite the limited sequence identity (<20%). The structure of thiamin phosphate synthase and indole-3-glycerophosphate synthase, which is a representative of the ribulose-phosphate binding barrel superfamily, are shown in Figure 4.15a. The RPS-BLAST alignment (E-value $1x10^{-10}$) (Figure 4.15a) and the DaliLite alignment (Z-score 15.4) of these two proteins are similar: 101 pairs of residues (~40% of the proteins) are equivalently aligned by the two comparison tools. Furthermore, three phosphate-binding residues are in equivalent positions both spatially and in the sequences of these proteins (Figure 4.15a). The homology between these two superfamilies has been previously reported (Nagano, Orengo et al. 2002).

The C-terminal domain of RNA polymerase alpha subunit and the DNA repair protein Rad51, N-terminal domain superfamilies are another pair of homologous

Figure 4.15: Homologous SCOP Superfamilies Identified by SCOPmap

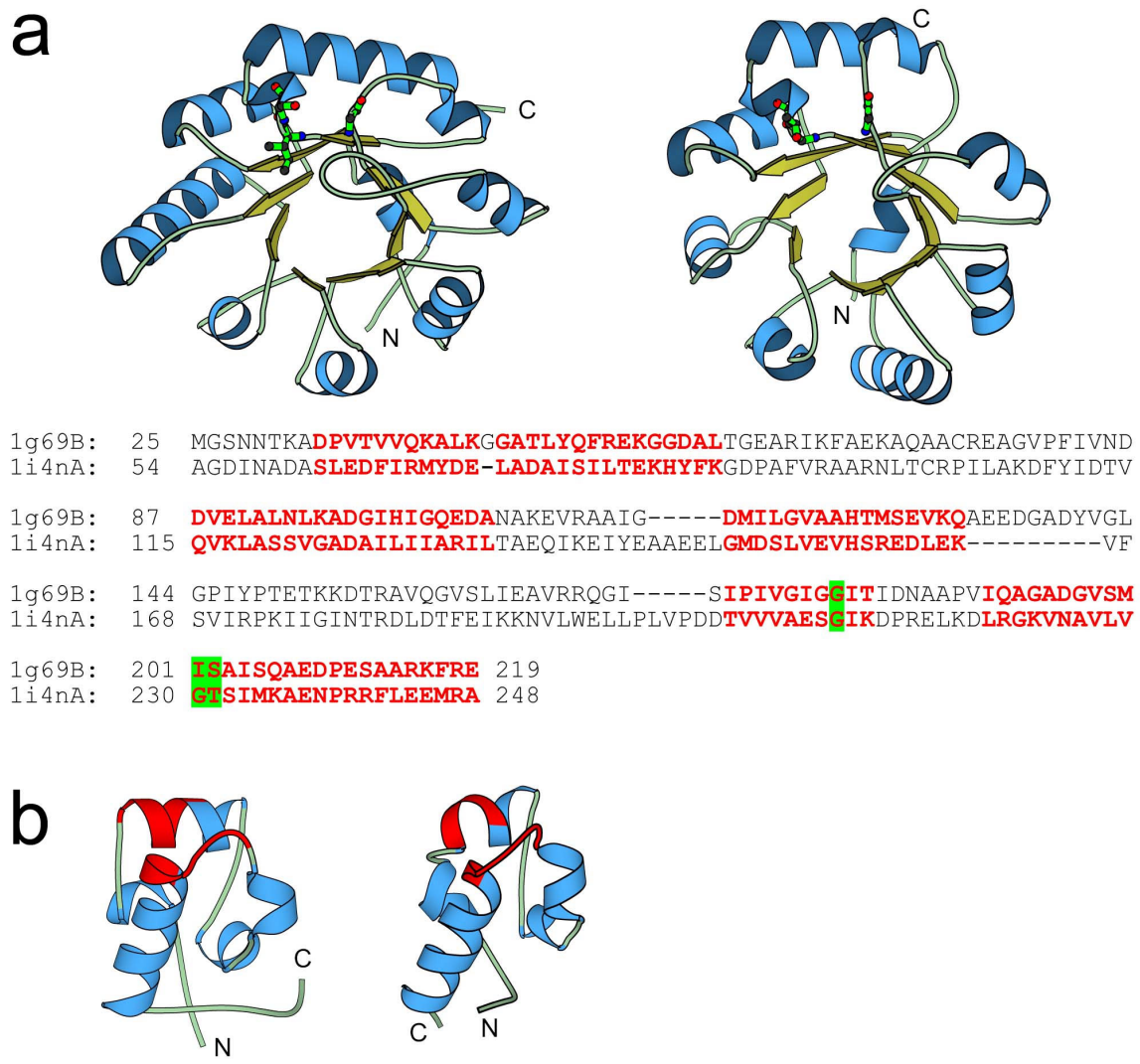


Figure 4.15: Homologous SCOP Superfamilies Identified by SCOPmap. a) MOLSCRIPT diagrams of thiamin phosphate synthase from *Bacillus subtilis* (left, pdb|1g69) from the thiamin phosphate synthase superfamily, and indole-3-glycerophosphate synthase from *Thermotoga maritima* (right, pdb|1i4n) from the ribulose-phosphate binding barrel superfamily. Pairwise alignment of representatives of these two superfamilies produced by PSI-BLAST. Residue pairs that are equivalently aligned by PSI-BLAST and DaliLite are showed in red bold letters. Numbers at the beginning and end of the alignment indicate the sequential residue number rather than the residue name assigned by the PDB file. Phosphate-binding residues in conserved positions in these two proteins are highlighted shown in green. b) MOLSCRIPT diagrams of α subunit C-terminal domain from *E. coli* RNA polymerase (left, pdb|1lb2) from the "C-terminal domain of RNA polymerase alpha subunit" superfamily, and the N-terminal domain of Rad51 from *Homo sapiens* (right, pdb|1b22) from the "DNA repair protein Rad51, N-terminal domain superfamily". Putative DNA-binding surfaces are shown in red.

superfamilies identified by SCOPmap. The domains in these two superfamilies have a 5-helix bundle structure (SAM domain-like fold), with one classic and one pseudo HhH motif as noted in SCOP. Members of both superfamilies have DNA-binding functions, and the observed or predicted DNA-binding surfaces are similar between the two superfamilies (Figure 4.15b). The closest representatives from each of these two superfamilies share ~32% sequence identity with each other. When the C-terminal domain of RNA polymerase alpha subunit superfamily is used as the query, all three members of this superfamily find hits to the single member of the DNA repair protein Rad51, N-terminal domain superfamily. These hits are identified by three different methods: RPS-BLAST (E-value 0.002), COMPASS (E-values ~10^{-16}), and MAMMOTH (Z-scores ~9). The detection of both confident sequence and structure comparison hits further supports the link between these two superfamilies.

4.5 PROGRAM AVAILABILITY

The SCOPmap script and instructions for library construction are available for download at ftp://iole.swmed.edu/pub/scopmap. SCOPmap results for representative PDB structures that are not included in the SCOP database are available at this site as well.

4.6 CONCLUSIONS

An algorithm, named SCOPmap, has been developed to assign domains in newly solved structures to appropriate superfamilies. The primary use of this program is to map domains within protein structures to an existing classification scheme. When applied to the SCOP database, this algorithm performs with ~95% accuracy (i.e. the correct superfamily assignment is made or no superfamily level assignment is made, as

appropriate). Comparison to SUPERFAMILY (an existing tool that performs a similar task of mapping queries to the SCOP database) shows that SCOPmap produces better results than SUPERFAMILY, both in terms of overall correct assignments and in the definition of the domain boundaries of those assignments. Examination of difficult cases has demonstrated the ability of SCOPmap to make non-trivial assignments, including some domains that represent common problems associated with automatic comparison tools. Although SCOPmap was developed in order to automatically assign proteins to SCOP, the utility of this algorithm is more general as the program could be modified to make assignments to other existing classification schemes as well.

CHAPTER 5:
Summary and Future Directions

5.1 CONCLUDING REMARKS: KINASE CLASSIFICATION

5.1.1 Project Summary

The comprehensive classification of available kinases (~59,000 sequences and 700 structures) is presented in Chapter 2. In this classification, the kinases are organized into fold groups, which reflect structural similarity. Within each fold group, the kinases are assembled into families of homologs. The structural fold, mode of nucleotide binding, and putative catalytic mechanism of each family identified in this work are described. Furthermore, the construction of this classification allows for investigation into how unrelated kinases carry out a similar biochemical reaction.

5.1.2 Applications and Utility

This classification should be beneficial to both experimental and computational biologists. First, this two-tier hierarchy allows for the structural and the evolutionary neighbors of a kinase-of-interest to be readily identified. Thus, potential uses for this classification include deduction of protein function (specifically, the phosphate-accepting substrate), inference of the nucleotide binding mode, or speculation on the enzymatic mechanism of poorly studied or newly discovered kinases based on proteins in the same family. Additionally, as a result of this work, structural annotations are now available for all known kinases, and the fold groups described in this classification present the final global picture of this entire functional class of proteins. This classification can also be used as the basis for more detailed studies of the individual kinase families, or in the investigation of other aspects of kinase function and evolution that were not specifically addressed in this work (e.g. mechanisms of kinase inhibition, determinants of substrate

specificity, the divergent evolution of similar or dissimilar kinase activities within families, etc). The insight gained from this analysis furthers the understanding of protein structure-function relationships in general and the evolution of various kinases in particular.

5.2 CONCLUDING REMARKS: SMALL DISULFIDE-RICH PROTEIN CLASSIFICATION

5.2.1 Project Summary

A comprehensive classification of available disulfide-rich protein domain structures is presented in Chapter 3. The disulfide-rich domains are organized into fold groups, which reflect structural similarity. Within each fold group, these domains are assembled into families of homologs. This work illustrates the functional and structural diversity among small disulfide-stabilized proteins. The classification reveals numerous examples of functional convergence among unrelated disulfide-rich proteins and functional divergence of homologous disulfide-rich domains. Variations in disulfide-bonding patterns among members of the same fold group or, in some cases, the same family are also examined.

5.2.2 Applications and Utility

The two-tier hierarchy of this classification should assist in the identification of structural and evolutionary neighbors of newly discovered disulfide-rich domains. This work should also be helpful in the study of known disulfide-rich domains, as the fold groups describe more broad similarities and the families comprise more remote evolutionary links than any other existing database addressing this structural class. A classification scheme recognizing distant relationships is especially useful for this

particular structural class, since links between members are often quite difficult to detect due to the nature of the small disulfide-rich domains themselves. Additionally, because the fold groups presented in this work signify cases of structural convergence, a thorough examination of these domains may give some insight into the debate concerning the underlying physical principles that govern the stability of disulfide-rich proteins. Thus, this classification should be useful for studying the evolution of the folds and functions of disulfide-rich domains in general, as well as for investigating the structural and evolutionary relatedness of specific disulfide-rich proteins in particular.

5.3 CONCLUDING REMARKS: SCOPMAP ALGORITHM FOR MAPPING PROTEIN DOMAINS TO AN EXISTING CLASSIFICATION

5.3.1 Project Summary

Chapter 4 describes the development of an algorithm (SCOPmap) designed to assign domains in newly solved protein structures to an existing classification database. The primary goal of SCOPmap is to identify homologs of the query structure, although structural neighbors are suggested in cases for which no homologous structural representatives are detected. The algorithm has been applied to the SCOP database, so that detection of homologs results in assignments of query domains at the SCOP superfamily level, while identification of structural neighbors results in assignments of query domains at the SCOP fold level.

5.3.2 Applications and Utility

This program should be of use to researchers interested in determining the evolutionary and structural neighbors of domains within newly solved protein structures. The SCOPmap algorithm can also suggest new evolutionary links between presumably

unrelated SCOP superfamilies (Kinch, Cheek et al. 2005). A modified version of the algorithm that utilizes only sequence comparison tools can be employed for the purposes of identifying remote homologs or making structural fold predictions for protein sequence queries. As the algorithm can run and subsequently compile the results from several different methods, this provides users with a single (and simple) tool for the analysis of large sets of sequence data with a variety of different well-known programs (gapped BLAST, PSI-BLAST, RPS-BLAST, and COMPASS).

Although SCOPmap was developed to automatically reproduce assignments to the SCOP classification, the strategy of this algorithm is more general and could be applied to any other related database as well. The algorithm could also potentially be used as an internal check in the preparation of new classifications or the maintenance and updating of existing classifications. Moreover, reliable methods for automatic updates to existing classification schemes become increasingly important with the rapid growth in sequence and structure databases.

BIBLIOGRAPHY

Adler, M, Lazarus, RA, Dennis, MS and Wagner, G. (1991) "Solution structure of kistrin, a potent platelet aggregation inhibitor and GP IIb-IIIa antagonist." *Science* **253**(5018):445-8.

Ahn, K and Kornberg, A. (1990) "Polyphosphate kinase from *Escherichia coli*. Purification and demonstration of a phosphoenzyme intermediate." *J Biol Chem* **265**(20):11734-9.

Aleshin, AE, Kirby, C, Liu, X, Bourenkov, GP, Bartunik, HD, Fromm, HJ and Honzatko, RB. (2000) "Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites and conformational changes relevant to allosteric regulation." *J Mol Biol* **296**(4):1001-15.

Alexander, JM, Nelson, CA, van Berkel, V, Lau, EK, Studts, JM, Brett, TJ, Speck, SH, Handel, TM, Virgin, HW and Fremont, DH. (2002) "Structural basis of chemokine sequestration by a herpesvirus decoy receptor." *Cell* **111**(3):343-56.

Allen, GS, Steinhauer, K, Hillen, W, Stulke, J and Brennan, RG. (2003) "Crystal structure of HPr kinase/phosphatase from *Mycoplasma pneumoniae*." *J Mol Biol* **326**(4):1203-17.

Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W and Lipman, DJ. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* **25**(17):3389-402.

Anand, R, Hoskins, AA, Stubbe, J and Ealick, SE. (2004) "Domain organization of *Salmonella typhimurium* formylglycinamide ribonucleotide amidotransferase revealed by X-ray crystallography." *Biochemistry* **43**(32):10328-42.

Anfinsen, CB and Scheraga, HA. (1975) "Experimental and theoretical aspects of protein folding." *Adv Protein Chem* **29**:205-300.

Apweiler, R, Attwood, TK, Bairoch, A, Bateman, A, Birney, E, Biswas, M, Bucher, P, Cerutti, L, Corpet, F, Croning, MD, et al. (2000) "InterPro--an integrated documentation resource for protein families, domains and functional sites." *Bioinformatics* **16**(12):1145-50.

Apweiler, R, Bairoch, A, Wu, CH, Barker, WC, Boeckmann, B, Ferro, S, Gasteiger, E, Huang, H, Lopez, R, Magrane, M, et al. (2004) "UniProt: the Universal Protein knowledgebase." *Nucleic Acids Res* **32**(Database issue):D115-9.

Arora, KK, Filburn, CR and Pedersen, PL. (1991) "Glucose phosphorylation. Site-directed mutations which impair the catalytic function of hexokinase." *J Biol Chem* **266**(9):5359-62.

Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT, et al. (2000) "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* **25**(1):25-9.

Aslund, F and Beckwith, J. (1999) "Bridge over troubled waters: sensing stress by disulfide bond formation." *Cell* **96**(6):751-3.

Attwood, TK, Beck, ME, Bleasby, AJ and Parry-Smith, DJ. (1994) "PRINTS--a database of protein motif fingerprints." *Nucleic Acids Res* **22**(17):3590-6.

Baker, LJ, Dorocke, JA, Harris, RA and Timm, DE. (2001) "The crystal structure of yeast thiamin pyrophosphokinase." *Structure* **9**(6):539-46.

Balaji, S, Sujatha, S, Kumar, SS and Srinivasan, N. (2001) "PALI-a database of Phylogeny and ALIgnment of homologous protein structures." *Nucleic Acids Res* **29**(1):61-5.

Barnham, KJ, Torres, AM, Alewood, D, Alewood, PF, Domagala, T, Nice, EC and Norton, RS. (1998) "Role of the 6-20 disulfide bridge in the structure and activity of epidermal growth factor." *Protein Sci* **7**(8):1738-49.

Barrett, AJ, Canter, CR, Liebecq, C, Moss, GP, Saenger, W, Sharon, N, Tipton, KF, Vnetianer, P and Vliegenthart, VFG. (1992). *Enzyme Nomenclature* Academic Press: San Diego, CA.

Barthe, P, Yang, YS, Chiche, L, Hoh, F, Strub, MP, Guignard, L, Soulier, J, Stern, MH, van Tilbeurgh, H, Lhoste, JM, et al. (1997) "Solution structure of human $p8^{MTCP1}$, a cysteine-rich protein encoded by the MTCP1 oncogene, reveals a new α-helical assembly motif." *J Mol Biol* **274**(5):801-15.

Bateman, A, Birney, E, Durbin, R, Eddy, SR, Howe, KL and Sonnhammer, EL. (2000) "The Pfam protein families database." *Nucleic Acids Res* **28**(1):263-6.

Bateman, A, Coin, L, Durbin, R, Finn, RD, Hollich, V, Griffiths-Jones, S, Khanna, A, Marshall, M, Moxon, S, Sonnhammer, EL, et al. (2004) "The Pfam protein families database." *Nucleic Acids Res* **32 Database issue**:D138-41.

Bauer, S, Kemter, K, Bacher, A, Huber, R, Fischer, M and Steinbacher, S. (2003) "Crystal structure of *Schizosaccharomyces pombe* riboflavin kinase reveals a novel ATP and riboflavin-binding fold." *J Mol Biol* **326**(5):1463-73.

Bayrhuber, M, Vijayan, V, Ferber, M, Graf, R, Korukottu, J, Imperial, J, Garrett, JE, Olivera, BM, Terlau, H, Zweckstetter, M, et al. (2005) "Conkunitzin-S1 is the first member of a new Kunitz-type neurotoxin family." *J Biol Chem* **280**(25):23766-70.

Benham, CJ and Jafri, MS. (1993) "Disulfide bonding patterns and protein topologies." *Protein Sci* **2**(1):41-54.

Berman, HM, Westbrook, J, Feng, Z, Gilliland, G, Bhat, TN, Weissig, H, Shindyalov, IN and Bourne, PE. (2000) "The Protein Data Bank." *Nucleic Acids Res* **28**(1):235-42.

Bernstein, BE and Hol, WG. (1998) "Crystal structures of substrates and products bound to the phosphoglycerate kinase active site reveal the catalytic mechanism." *Biochemistry* **37**(13):4429-36.

Bernstein, BE, Michels, PA and Hol, WG. (1997) "Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation." *Nature* **385**(6613):275-8.

Betz, SF. (1993) "Disulfide bonds and the stability of globular proteins." *Protein Sci* **2**(10):1551-8.

Bilgrami, S, Tomar, S, Yadav, S, Kaur, P, Kumar, J, Jabeen, T, Sharma, S and Singh, TP. (2004) "Crystal structure of schistatin, a disintegrin homodimer from saw-scaled viper (*Echis carinatus*) at 2.5 Å resolution." *J Mol Biol* **341**(3):829-37.

Bilgrami, S, Yadav, S, Kaur, P, Sharma, S, Perbandt, M, Betzel, C and Singh, TP. (2005) "Crystal Structure of the Disintegrin Heterodimer from Saw-Scaled Viper (*Echis carinatus*) at 1.9 Å Resolution." *Biochemistry* **44**(33):11058-66.

Bilwes, AM, Quezada, CM, Croal, LR, Crane, BR and Simon, MI. (2001) "Nucleotide binding by the histidine kinase CheA." *Nat Struct Biol* **8**(4):353-60.

Blaszczyk, J, Shi, G, Yan, H and Ji, X. (2000) "Catalytic center assembly of HPPK as revealed by the crystal structure of a ternary complex at 1.25 Å resolution." *Structure* **8**(10):1049-58.

Blom, NS, Tetreault, S, Coulombe, R and Sygusch, J. (1996) "Novel active site in *Escherichia coli* fructose 1,6-bisphosphate aldolase." *Nat Struct Biol* **3**(10):856-62.

Bode, W, Engh, R, Musil, D, Thiele, U, Huber, R, Karshikov, A, Brzin, J, Kos, J and Turk, V. (1988) "The 2.0 Å X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases." *EMBO J* **7**(8):2593-9.

Borden, KL and Freemont, PS. (1996) "The RING finger domain: a recent example of a sequence-structure family." *Curr Opin Struct Biol* **6**(3):395-401.

Bork, P, Sander, C and Valencia, A. (1992) "An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins." *Proc Natl Acad Sci U S A* **89**(16):7290-4.

Bork, P, Sander, C and Valencia, A. (1993) "Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases." *Protein Sci* **2**(1):31-40.

Bossemeyer, D, Engh, RA, Kinzel, V, Ponstingl, H and Huber, R. (1993) "Phosphotransferase and substrate binding mechanism of the cAMP-dependent protein kinase catalytic subunit from porcine heart as deduced from the 2.0 Å structure of the complex with Mn^{2+} adenylyl imidodiphosphate and inhibitor peptide PKI(5-24)." *EMBO J* **12**(3):849-59.

Bray, JE, Todd, AE, Pearl, FM, Thornton, JM and Orengo, CA. (2000) "The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues." *Protein Eng* **13**(3):153-65.

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

Brinkkotter, A, Kloss, H, Alpert, C and Lengeler, JW. (2000) "Pathways for the utilization of N-acetyl-galactosamine and galactosamine in *Escherichia coli*." *Mol Microbiol* **37**(1):125-35.

Brown, LR, Mronga, S, Bradshaw, RA, Ortenzi, C, Luporini, P and Wuthrich, K. (1993) "Nuclear magnetic resonance solution structure of the pheromone ER-10 from the ciliated protozoan *Euplotes raikovi*." *J Mol Biol* **231**(3):800-16.

Brown, RL, Haley, TL, West, KA and Crabb, JW. (1999) "Pseudechetoxin: a peptide blocker of cyclic nucleotide-gated ion channels." *Proc Natl Acad Sci U S A* **96**(2):754-9.

Brun, C, Chevenet, F, Martin, D, Wojcik, J, Guenoche, A and Jacq, B. (2003) "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network." *Genome Biol* **5**(1):R6.

Buchan, DW, Rison, SC, Bray, JE, Lee, D, Pearl, F, Thornton, JM and Orengo, CA. (2003) "Gene3D: structural assignments for the biologist and bioinformaticist alike." *Nucleic Acids Res* **31**(1):469-73.

Bujnicki, JM, Elofsson, A, Fischer, D and Rychlewski, L. (2001) "Structure prediction meta server." *Bioinformatics* **17**(8):750-1.

Burk, DL, Hon, WC, Leung, AK and Berghuis, AM. (2001) "Structural analyses of nucleotide binding to an aminoglycoside phosphotransferase." *Biochemistry* **40**(30):8756-64.

Buss, KA, Cooper, DR, Ingram-Smith, C, Ferry, JG, Sanders, DA and Hasson, MS. (2001) "Urkinase: structure of acetate kinase, a member of the ASKHA superfamily of phosphotransferases." *J Bacteriol* **183**(2):680-6.

Buss, KA, Ingram-Smith, C, Ferry, JG, Sanders, DA and Hasson, MS. (1997) "Crystallization of acetate kinase from *Methanosarcina thermophila* and prediction of its fold." *Protein Sci* **6**(12):2659-62.

Calder, RB, Williams, RS, Ramaswamy, G, Rock, CO, Campbell, E, Unkles, SE, Kinghorn, JR and Jackowski, S. (1999) "Cloning and characterization of a eukaryotic pantothenate kinase gene (panK) from *Aspergillus nidulans*." *J Biol Chem* **274**(4):2014-20.

Calvete, JJ, Marcinkiewicz, C, Monleon, D, Esteve, V, Celda, B, Juarez, P and Sanz, L. (2005) "Snake venom disintegrins: evolution of structure and function." *Toxicon* **45**(8):1063-74.

Campos-Olivas, R, Bruix, M, Santoro, J, Lacadena, J, Martinez del Pozo, A, Gavilanes, JG and Rico, M. (1995) "NMR solution structure of the antifungal protein from *Aspergillus giganteus*: evidence for cysteine pairing isomerism." *Biochemistry* **34**(9):3009-21.

Carret, C, Delbecq, S, Labesse, G, Carcy, B, Precigout, E, Moubri, K, Schetters, TP and Gorenflot, A. (1999) "Characterization and molecular cloning of an adenosine kinase from *Babesia canis rossi*." *Eur J Biochem* **265**(3):1015-21.

Chang, C, Quartey, P, Shiu, M, Collart, F and Joachimiak, A. (*To be published*) "Crystal Structure of Hypothetical Protein from *Porphyromonas gingivalis*."

Chistoserdova, L and Lidstrom, ME. (1997) "Identification and mutation of a gene required for glycerate kinase activity from a facultative methylotroph, *Methylobacterium extorquens* AM1." *J Bacteriol* **179**(15):4946-8.

Cho, H, Wang, W, Kim, R, Yokota, H, Damo, S, Kim, SH, Wemmer, D, Kustu, S and Yan, D. (2001) "BeF_3^- acts as a phosphate analog in proteins phosphorylated on aspartate: structure of a BeF_3^- complex with phosphoserine phosphatase." *Proc Natl Acad Sci U S A* **98**(15):8525-30.

Cho, HS and Leahy, DJ. (2002) "Structure of the extracellular region of HER3 reveals an interdomain tether." *Science* **297**(5585):1330-3.

Collet, JF, Stroobant, V and Van Schaftingen, E. (1999) "Mechanistic studies of phosphoserine phosphatase, an enzyme related to P-type ATPases." *J Biol Chem* **274**(48):33985-90.

Craik, DJ, Simonsen, S and Daly, NL. (2002) "The cyclotides: novel macrocyclic peptides as scaffolds in drug design." *Curr Opin Drug Discov Devel* **5**(2):251-60.

Creighton, TE. (1992) "Protein folding pathways determined using disulphide bonds." *Bioessays* **14**(3):195-9.

Creighton, TE. (1997) "Protein folding coupled to disulphide bond formation." *Biol Chem* **378**(8):731-44.

Crouzet, P and Otten, L. (1995) "Sequence and mutational analysis of a tartrate utilization operon from *Agrobacterium vitis*." *J Bacteriol* **177**(22):6518-26.

Cuff, JA, Clamp, ME, Siddiqui, AS, Finlay, M and Barton, GJ. (1998) "JPred: a consensus secondary structure prediction server." *Bioinformatics* **14**(10):892-3.

Daly, NL, Clark, RJ and Craik, DJ. (2003) "Disulfide folding pathways of cystine knot proteins." *J Biol Chem* **278**(8):6314-22.

Daugherty, M, Vonstein, V, Overbeek, R and Osterman, A. (2001) "Archaeal shikimate kinase, a new member of the GHMP-kinase family." *J Bacteriol* **183**(1):292-300.

Dauplais, M, Lecoq, A, Song, J, Cotton, J, Jamin, N, Gilquin, B, Roumestand, C, Vita, C, de Medeiros, CL, Rowan, EG, et al. (1997) "On the convergent evolution of animal toxins. Conservation of a diad of functional residues in potassium channel-blocking toxins with unrelated structures." *J Biol Chem* **272**(7):4302-9.

Davies, DR, Interthal, H, Champoux, JJ and Hol, WG. (2002) "The crystal structure of human tyrosyl-DNA phosphodiesterase, Tdp1." *Structure (Camb)* **10**(2):237-48.

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

Doig, AJ and Williams, DH. (1991) "Is the hydrophobic effect stabilizing or destabilizing in proteins? The contribution of disulphide bonds to protein stability." *J Mol Biol* **217**(2):389-98.

Donald, JE and Shakhnovich, EI. (2005) "Predicting specificity-determining residues in two large eukaryotic transcription factor families." *Nucleic Acids Res* **33**(14):4455-65.

Dutta, R, Qin, L and Inouye, M. (1999) "Histidine kinases: diversity of domain organization." *Mol Microbiol* **34**(4):633-40.

Eddy, SR. (1998) "Profile hidden Markov models." *Bioinformatics* **14**(9):755-63.

Eigenbrot, C, Randal, M and Kossiakoff, AA. (1990) "Structural effects induced by removal of a disulfide-bridge: the X-ray structure of the C30A/C51A mutant of basic pancreatic trypsin inhibitor at 1.6 Å." *Protein Eng* **3**(7):591-8.

Escoubas, P, Diochot, S and Corzo, G. (2000) "Structure and pharmacology of spider venom neurotoxins." *Biochimie* **82**(9-10):893-907.

Espiritu, DJ, Watkins, M, Dia-Monje, V, Cartier, GE, Cruz, LJ and Olivera, BM. (2001) "Venomous cone snails: molecular phylogeny and the generation of toxin diversity." *Toxicon* **39**(12):1899-916.

Eswaramoorthy, S and Swaminathan, S. (*To be published*) "Crystal Structure of Thiamine Monophosphate Kinase (ThiL) from *Aquifex aeolicus*."

Evans, PR, Farrants, GW and Hudson, PJ. (1981) "Phosphofructokinase: structure and control." *Philos Trans R Soc Lond B Biol Sci* **293**(1063):53-62.

Fanutti, C, Ponyi, T, Black, GW, Hazlewood, GP and Gilbert, HJ. (1995) "The conserved noncatalytic 40-residue sequence in cellulases and hemicellulases from anaerobic fungi functions as a protein docking domain." *J Biol Chem* **270**(49):29314-22.

Fieulaine, S, Morera, S, Poncet, S, Monedero, V, Gueguen-Chaignon, V, Galinier, A, Janin, J, Deutscher, J and Nessler, S. (2001) "X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain." *EMBO J* **20**(15):3917-27.

Fischer, D. (2000) "Hybrid fold recognition: combining sequence derived properties with evolutionary information." *Pac Symp Biocomput*:119-30.

Flory, PJ. (1956) "Theory of elastic mechanisms in fibrous proteins." *J Am Chem Soc* **78**:5222-35.

Fritz-Wolf, K, Schnyder, T, Wallimann, T and Kabsch, W. (1996) "Structure of mitochondrial creatine kinase." *Nature* **381**(6580):341-5.

Galperin, MY, Walker, DR and Koonin, EV. (1998) "Analogous enzymes: independent inventions in enzyme evolution." *Genome Res* **8**(8):779-90.

Garman, SC, Simcoke, WN, Stowers, AW and Garboczi, DN. (2003) "Structure of the C-terminal domains of merozoite surface protein-1 from *Plasmodium knowlesi* reveals a novel histidine binding site." *J Biol Chem* **278**(9):7264-9.

Gasparini, S, Danse, JM, Lecoq, A, Pinkasfeld, S, Zinn-Justin, S, Young, LC, de Medeiros, CC, Rowan, EG, Harvey, AL and Menez, A. (1998) "Delineation of the functional site of α-dendrotoxin." *J Biol Chem* **273**(39):25393-403.

Gattiker, A, Gasteiger, E and Bairoch, A. (2002) "ScanProsite: a reference implementation of a PROSITE scanning tool." *Appl Bioinformatics* **1**(2):107-8.

Geer, LY, Domrachev, M, Lipman, DJ and Bryant, SH. (2002) "CDART: protein homology by domain architecture." *Genome Res* **12**(10):1619-23.

Gelly, JC, Gracy, J, Kaas, Q, Le-Nguyen, D, Heitz, A and Chiche, L. (2004) "The KNOTTIN website and database: a new information system dedicated to the knottin scaffold." *Nucleic Acids Res* **32**(Database issue):D156-9.

Georgiadis, MM, Jessen, SM, Ogata, CM, Telesnitsky, A, Goff, SP and Hendrickson, WA. (1995) "Mechanistic implications from the structure of a catalytic fragment of Moloney murine leukemia virus reverse transcriptase." *Structure* **3**(9):879-92.

Gerdes, SY, Scholle, MD, D'Souza, M, Bernal, A, Baev, MV, Farrell, M, Kurnasov, OV, Daugherty, MD, Mseeh, F, Polanuyer, BM, et al. (2002) "From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways." *J Bacteriol* **184**(16):4555-72.

Getz, G, Vendruscolo, M, Sachs, D and Domany, E. (2002) "Automated assignment of SCOP and CATH protein structure classifications from FSSP scores." *Proteins* **46**(4):405-15.

Gil-Ortiz, F, Ramon-Maiques, S, Fita, I and Rubio, V. (2003) "The course of phosphorus in the reaction of N-acetyl-L-glutamate kinase, determined from the structures of crystalline complexes, including a complex with an AlF_4^- transition state mimic." *J Mol Biol* **331**(1):231-44.

Ginalski, K, Elofsson, A, Fischer, D and Rychlewski, L. (2003) "3D-Jury: a simple approach to improve protein structure predictions." *Bioinformatics* **19**(8):1015-8.

Ginalski, K, Pas, J, Wyrwicz, LS, von Grotthuss, M, Bujnicki, JM and Rychlewski, L. (2003) "ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure." *Nucleic Acids Res* **31**(13):3804-7.

Ginalski, K and Rychlewski, L. (2003) "Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment." *Proteins* **53 Suppl 6**:410-7.

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

Goldsmith, EJ and Cobb, MH. (1994) "Protein kinases." *Curr Opin Struct Biol* **4**(6):833-40.

Gonzalez, B, Schell, MJ, Letcher, AJ, Veprintsev, DB, Irvine, RF and Williams, RL. (2004) "Structure of a human inositol 1,4,5-trisphosphate 3-kinase: substrate binding reveals why it is not a phosphoinositide 3-kinase." *Mol Cell* **15**(5):689-701.

Goss, NH, Evans, CT and Wood, HG. (1980) "Pyruvate phosphate dikinase: sequence of the histidyl peptide, the pyrophosphoryl and phosphoryl carrier." *Biochemistry* **19**(25):5805-9.

Gough, J, Karplus, K, Hughey, R and Chothia, C. (2001) "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure." *J Mol Biol* **313**(4):903-19.

Gould, RJ, Polokoff, MA, Friedman, PA, Huang, TF, Holt, JC, Cook, JJ and Niewiarowski, S. (1990) "Disintegrins: a family of integrin inhibitory proteins from viper venoms." *Proc Soc Exp Biol Med* **195**(2):168-71.

Grishin, NV. (1999) "Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds." *J Mol Biol* **291**(2):239-47.

Guo, M, Teng, M, Niu, L, Liu, Q, Huang, Q and Hao, Q. (2005) "Crystal structure of the cysteine-rich secretory protein stecrisp reveals that the cysteine-rich domain has a K^+ channel inhibitor-like fold." *J Biol Chem* **280**(13):12405-12.

Gupta, A, Van Vlijmen, HW and Singh, J. (2004) "A classification of disulfide patterns and its relationship to protein structure and function." *Protein Sci* **13**(8):2045-58.

Haft, DH, Loftus, BJ, Richardson, DL, Yang, F, Eisen, JA, Paulsen, IT and White, O. (2001) "TIGRFAMs: a protein family resource for the functional identification of proteins." *Nucleic Acids Res* **29**(1):41-3.

Hall, DR, Bond, CS, Leonard, GA, Watt, CI, Berry, A and Hunter, WN. (2002) "Structure of tagatose-1,6-bisphosphate aldolase." *J Biol Chem* **277**(24):22018-24.

Hanks, SK and Hunter, T. (1995) "Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification." *FASEB Journal* **9**(8):576-96.

Harrison, A, Pearl, F, Sillitoe, I, Slidel, T, Mott, R, Thornton, J and Orengo, C. (2003) "Recognizing the fold of a protein structure." *Bioinformatics* **19**(14):1748-59.

Harrison, PM and Sternberg, MJ. (1996) "The disulphide β-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds." *J Mol Biol* **264**(3):603-23.

Hart, JC, Hillier, IH, Burton, NA and Sheppard, DW. (1998) "An Alternative Role for the Conserved Asp Residue in Phosphoryl Transfer Reactions." *J Am Chem Soc* **120**(51):13535-6.

Hartig, GR, Tran, TT and Smythe, ML. (2005) "Intramolecular disulphide bond arrangements in nonhomologous proteins." *Protein Sci* **14**(2):474-82.

Hatano, K, Kojima, M, Tanokura, M and Takahashi, K. (1995) "Primary structure, sequence-specific 1H-NMR assignments and secondary structure in solution of bromelain inhibitor VI from pineapple stem." *Eur J Biochem* **232**(2):335-43.

Hegyi, H and Gerstein, M. (1999) "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome." *J Mol Biol* **288**(1):147-64.

Heikinheimo, P, Helland, R, Leiros, HK, Leiros, I, Karlsen, S, Evjen, G, Ravelli, R, Schoehn, G, Ruigrok, R, Tollersrud, OK, et al. (2003) "The structure of bovine

lysosomal α-mannosidase suggests a novel mechanism for low-pH activation." *J Mol Biol* **327**(3):631-44.

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

Herzberg, O, Chen, CC, Kapadia, G, McGuire, M, Carroll, LJ, Noh, SJ and Dunaway-Mariano, D. (1996) "Swiveling-domain mechanism for enzymatic phosphotransfer between remote reaction sites." *Proc Natl Acad Sci U S A* **93**(7):2652-7.

Hewage, CM, Jiang, L, Parkinson, JA, Ramage, R and Sadler, IH. (1999) "Solution structure of a novel ET_B receptor selective agonist ET_{1-21} [$Cys(Acm)^{1,15}$, $Aib^{3,11}$, Leu^{7}] by nuclear magnetic resonance spectroscopy and molecular modelling." *J Pept Res* **53**(3):223-33.

Hisano, T, Hata, Y, Fujii, T, Liu, JQ, Kurihara, T, Esaki, N and Soda, K. (1996) "Crystal structure of L-2-haloacid dehalogenase from *Pseudomonas sp*. YL. An α/β hydrolase structure that is different from the α/β hydrolase fold." *J Biol Chem* **271**(34):20322-30.

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

Holm, L and Sander, C. (1994) "The FSSP database of structurally aligned protein fold families." *Nucleic Acids Res* **22**(17):3600-9.

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

Holm, L and Sander, C. (1997) "Decision support system for the evolutionary classification of protein structures." *Proc Int Conf Intell Syst Mol Biol* **5**:140-6.

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

Hondoh, H, Kuriki, T and Matsuura, Y. (2003) "Three-dimensional structure and substrate binding of *Bacillus stearothermophilus* neopullulanase." *J Mol Biol* **326**(1):177-88.

Horn, NA, Hurst, GB, Mayasundari, A, Whittemore, NA, Serpersu, EH and Peterson, CB. (2004) "Assignment of the four disulfides in the N-terminal somatomedin B domain of native vitronectin isolated from human plasma." *J Biol Chem* **279**(34):35867-78.

Huang, K, Strynadka, NC, Bernard, VD, Peanasky, RJ and James, MN. (1994) "The molecular structure of the complex of *Ascaris* chymotrypsin/elastase inhibitor with porcine elastase." *Structure* **2**(7):679-89.

Hutter, MC and Helms, V. (2000) "Phosphoryl transfer by a concerted reaction mechanism in UMP/CMP-kinase." *Protein Sci* **9**(11):2225-31.

Ito, S, Fushinobu, S, Yoshioka, I, Koga, S, Matsuzawa, H and Wakagi, T. (2001) "Structural basis for the ADP-specificity of a novel glucokinase from a hyperthermophilic archaeon." *Structure* **9**(3):205-14.

Janin, J, Dumas, C, Morera, S, Xu, Y, Meyer, P, Chiadmi, M and Cherfils, J. (2000) "Three-dimensional structure of nucleoside diphosphate kinase." *J Bioenerg Biomembr* **32**(3):215-25.

Jing, H, Takagi, J, Liu, JH, Lindgren, S, Zhang, RG, Joachimiak, A, Wang, JH and Springer, TA. (2002) "Archaeal surface layer proteins contain β propeller, PKD, and β helix domains and are related to metazoan cell surface proteins." *Structure (Camb)* **10**(10):1453-64.

Joint Center for Structural Genomics. (*To be published*) "Crystal Structure of Glycerate Kinase (Tm1585) from *Thermotoga maritima* at 2.95 Å Resolution."

Jomaa, H, Wiesner, J, Sanderbrand, S, Altincicek, B, Weidemeyer, C, Hintz, M, Turbachova, I, Eberl, M, Zeidler, J, Lichtenthaler, HK, et al. (1999) "Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs." *Science* **285**(5433):1573-6.

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

Kalus, W, Zweckstetter, M, Renner, C, Sanchez, Y, Georgescu, J, Grol, M, Demuth, D, Schumacher, R, Dony, C, Lang, K, et al. (1998) "Structure of the IGF-binding domain of the insulin-like growth factor-binding protein-5 (IGFBP-5): implications for IGF and IGF-I receptor interactions." *EMBO J* **17**(22):6558-72.

Kamikubo, Y, De Guzman, R, Kroon, G, Curriden, S, Neels, JG, Churchill, MJ, Dawson, P, Oldziej, S, Jagielska, A, Scheraga, HA, et al. (2004) "Disulfide bonding arrangements in active forms of the somatomedin B domain of human vitronectin." *Biochemistry* **43**(21):6519-34.

Karmirantzou, M and Hamodrakas, SJ. (2001) "A Web-based classification system of DNA-binding protein families." *Protein Eng* **14**(7):465-72.

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

Karthikeyan, S, Zhou, Q, Mseeh, F, Grishin, NV, Osterman, AL and Zhang, H. (2003a) "Crystal structure of human riboflavin kinase reveals a β barrel fold and a novel active site arch." *Structure (Camb)* **11**(3):265-73.

Karthikeyan, S, Zhou, Q, Osterman, AL and Zhang, H. (2003b) "Ligand binding-induced conformational changes in riboflavin kinase: structural basis for the ordered mechanism." *Biochemistry* **42**(43):12532-8.

Kelley, LA, MacCallum, RM and Sternberg, MJ. (2000) "Enhanced genome annotation using structural profiles in the program 3D-PSSM." *J Mol Biol* **299**(2):499-520.

Kim, CH and King, TE. (1983) "A mitochondrial protein essential for the formation of the cytochrome c1-c complex. Isolation, purification, and properties." *J Biol Chem* **258**(22):13543-51.

Kinch, LN, Cheek, S and Grishin, NV. (2005) "EDD, a novel phosphotransferase domain common to mannose transporter EIIA, dihydroxyacetone kinase, and DegV." *Protein Sci* **14**(2):360-7.

Kohno, T, Sasaki, T, Kobayashi, K, Fainzilber, M and Sato, K. (2002) "Three-dimensional solution structure of the sodium channel agonist/antagonist δ-conotoxin TxVIA." *J Biol Chem* **277**(39):36387-91.

Korolev, SV, Dementieva, IS, Christendat, D, Edwards, A and Joachimiak, A. (*To be published*) "Structural Similarities of Mth1747 Hypothetical Protein from *Methanobacterium thermoautotrophicum* with 3-Hydroxyacid Dehydrogenases."

Kozlov, G, Perreault, A, Schrag, JD, Park, M, Cygler, M, Gehring, K and Ekiel, I. (2004) "Insights into function of PSI domains from structure of the Met receptor PSI domain." *Biochem Biophys Res Commun* **321**(1):234-40.

Kraulis, PJ. (1991) "MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures." *J of App Crystall* **24**(5):946-950.

Krishna, SS, Majumdar, I and Grishin, NV. (2003) "Structural classification of zinc fingers: survey and summary." *Nucleic Acids Res* **31**(2):532-50.

Krishna, SS, Zhou, T, Daugherty, M, Osterman, A and Zhang, H. (2001) "Structural basis for the catalysis and substrate specificity of homoserine kinase." *Biochemistry* **40**(36):10810-8.

Krissinel, E and Henrick, K (2003). Protein structure comparison in 3D based on secondary structure matching (SSM) followed by C_α alignment, scored by a new structural similarity function. Proceedings of the 5th International Conference on Molecular Structural Biology, Vienna.

Krissinel, E and Henrick, K. (2004) "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions." *Acta Crystallogr D Biol Crystallogr* **60**(Pt 12 Pt 1):2256-68.

Kryshtafovych, A, Venclovas, C, Fidelis, K and Moult, J. (2005) "Progress over the first decade of CASP experiments." *Proteins*.

Kuettner, EB, Hilgenfeld, R and Weiss, MS. (2002a) "The active principle of garlic at atomic resolution." *J Biol Chem* **277**(48):46402-7.

Kuettner, EB, Hilgenfeld, R and Weiss, MS. (2002b) "Purification, characterization, and crystallization of alliinase from garlic." *Arch Biochem Biophys* **402**(2):192-200.

Kumble, KD, Ahn, K and Kornberg, A. (1996) "Phosphohistidyl active sites in polyphosphate kinase of *Escherichia coli*." *Proc Natl Acad Sci U S A* **93**(25):14391-5.

Kuroda, A and Kornberg, A. (1997) "Polyphosphate kinase as a nucleoside diphosphate kinase in *Escherichia coli* and *Pseudomonas aeruginosa*." *Proc Natl Acad Sci U S A* **94**(2):439-42.

Lamoureux, JS, Stuart, D, Tsang, R, Wu, C and Glover, JN. (2002) "Structure of the sporulation-specific transcription factor Ndt80 bound to DNA." *EMBO J* **21**(21):5721-32.

Larsen, TM, Benning, MM, Rayment, I and Reed, GH. (1998) "Structure of the bis(Mg^{2+})-ATP-oxalate complex of the rabbit muscle pyruvate kinase at 2.1 Å resolution: ATP binding over a barrel." *Biochemistry* **37**(18):6247-55.

Larsen, TM, Laughlin, LT, Holden, HM, Rayment, I and Reed, GH. (1994) "Structure of rabbit muscle pyruvate kinase complexed with Mn^{2+}, K^+, and pyruvate." *Biochemistry* **33**(20):6301-9.

Laskowski, M and Qasim, MA. (2000) "What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes?" *Biochim Biophys Acta* **1477**(1-2):324-37.

Lauber, T, Schulz, A, Schweimer, K, Adermann, K and Marx, UC. (2003) "Homologous proteins with different folds: the three-dimensional structures of domains 1 and 6 of the multiple Kazal-type inhibitor LEKTI." *J Mol Biol* **328**(1):205-19.

Lee, BI, Chang, C, Cho, SJ, Eom, SH, Kim, KK, Yu, YG and Suh, SW. (2001) "Crystal structure of the MJ0490 gene product of the hyperthermophilic archaebacterium *Methanococcus jannaschii*, a novel member of the lactate/malate family of dehydrogenases." *J Mol Biol* **307**(5):1351-62.

Leonard, CJ, Aravind, L and Koonin, EV. (1998) "Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily." *Genome Res* **8**(10):1038-47.

Letunic, I, Copley, RR, Schmidt, S, Ciccarelli, FD, Doerks, T, Schultz, J, Ponting, CP and Bork, P. (2004) "SMART 4.0: towards genomic data integration." *Nucleic Acids Res* **32 Database issue**:D142-4.

Li, C, Kappock, TJ, Stubbe, J, Weaver, TM and Ealick, SE. (1999) "X-ray crystal structure of aminoimidazole ribonucleotide synthetase (PurM), from the *Escherichia coli* purine biosynthetic pathway at 2.5 Å resolution." *Structure* **7**(9):1155-66.

Li, MH, Kwok, F, Chang, WR, Lau, CK, Zhang, JP, Lo, SC, Jiang, T and Liang, DC. (2002) "Crystal structure of brain pyridoxal kinase, a novel member of the ribokinase superfamily." *J Biol Chem* **277**(48):46385-90.

Lichtenthaler, HK, Zeidler, J, Schwender, J and Muller, C. (2000) "The non-mevalonate isoprenoid biosynthesis of plants as a test system for new herbicides and drugs against pathogenic bacteria and the malaria parasite." *Zeitschrift fur Naturforschung* **55**(5-6):305-13.

Locher, KP, Hans, M, Yeh, AP, Schmid, B, Buckel, W and Rees, DC. (2001) "Crystal structure of the *Acidaminococcus fermentans* 2-hydroxyglutaryl-CoA dehydratase component A." *J Mol Biol* **307**(1):297-308.

Loris, R, Marianovsky, I, Lah, J, Laeremans, T, Engelberg-Kulka, H, Glaser, G, Muyldermans, S and Wyns, L. (2003) "Crystal structure of the intrinsically flexible addiction antidote MazE." *J Biol Chem* **278**(30):28252-7.

Luttgen, H, Rohdich, F, Herz, S, Wungsintaweekul, J, Hecht, S, Schuhr, CA, Fellermeier, M, Sagner, S, Zenk, MH, Bacher, A, et al. (2000) "Biosynthesis of terpenoids: YchB protein of *Escherichia coli* phosphorylates the 2-hydroxy group of 4-diphosphocytidyl-2C-methyl-D-erythritol." *Proc Natl Acad Sci U S A* **97**(3):1062-7.

Machius, M, Chuang, JL, Wynn, RM, Tomchick, DR and Chuang, DT. (2001) "Structure of rat BCKD kinase: nucleotide-induced domain communication in a mitochondrial protein kinase." *Proc Natl Acad Sci U S A* **98**(20):11218-23.

Marchler-Bauer, A, Anderson, JB, Cherukuri, PF, DeWeese-Scott, C, Geer, LY, Gwadz, M, He, S, Hurwitz, DI, Jackson, JD, Ke, Z, et al. (2005) "CDD: a Conserved Domain Database for protein classification." *Nucleic Acids Res* **33**(Database issue):D192-6.

Marchler-Bauer, A, Anderson, JB, DeWeese-Scott, C, Fedorova, ND, Geer, LY, He, S, Hurwitz, DI, Jackson, JD, Jacobs, AR, Lanczycki, CJ, et al. (2003) "CDD: a curated Entrez database of conserved domain alignments." *Nucleic Acids Res* **31**(1):383-7.

Marina, A, Alzari, PM, Bravo, J, Uriarte, M, Barcelona, B, Fita, I and Rubio, V. (1999) "Carbamate kinase: New structural machinery for making carbamoyl phosphate, the common precursor of pyrimidines and arginine." *Protein Sci* **8**(4):934-40.

Mas, JM, Aloy, P, Marti-Renom, MA, Oliva, B, Blanco-Aparicio, C, Molina, MA, de Llorens, R, Querol, E and Aviles, FX. (1998) "Protein similarities beyond disulphide bridge topology." *J Mol Biol* **284**(3):541-8.

Mas, JM, Aloy, P, Marti-Renom, MA, Oliva, B, de Llorens, R, Aviles, FX and Querol, E. (2001) "Classification of protein disulphide-bridge topologies." *J Comput Aided Mol Des* **15**(5):477-87.

Matsuo, Y and Bryant, SH. (1999) "Identification of homologous core structures." *Proteins* **35**(1):70-9.

Matte, A, Goldie, H, Sweet, RM and Delbaere, LT. (1996) "Crystal structure of *Escherichia coli* phosphoenolpyruvate carboxykinase: a new structural family with the P-loop nucleoside triphosphate hydrolase fold." *J Mol Biol* **256**(1):126-43.

Matte, A, Tari, LW and Delbaere, LT. (1998) "How do kinases transfer phosphoryl groups?" *Structure* **6**(4):413-9.

Mayasundari, A, Whittemore, NA, Serpersu, EH and Peterson, CB. (2004) "The solution structure of the N-terminal domain of human vitronectin: proximal sites that regulate fibrinolysis and cell migration." *J Biol Chem* **279**(28):29359-66.

McManus, AM, Nielsen, KJ, Marcus, JP, Harrison, SJ, Green, JL, Manners, JM and Craik, DJ. (1999) "MiAMP1, a novel protein from *Macadamia integrifolia* adopts

a Greek key β-barrel fold unique amongst plant antimicrobial proteins." *J Mol Biol* **293**(3):629-38.

Meinhart, A, Alonso, JC, Strater, N and Saenger, W. (2003) "Crystal structure of the plasmid maintenance system ε/ζ: functional mechanism of toxin ζ and inactivation by $\varepsilon_2\zeta_2$ complex formation." *Proc Natl Acad Sci U S A* **100**(4):1661-6.

Menez, A. (1998) "Functional architectures of animal toxins: a clue to drug design?" *Toxicon* **36**(11):1557-72.

Miles, LA, Dy, CY, Nielsen, J, Barnham, KJ, Hinds, MG, Olivera, BM, Bulaj, G and Norton, RS. (2002) "Structure of a novel P-superfamily spasmodic conotoxin reveals an inhibitory cystine knot motif." *J Biol Chem* **277**(45):43033-40.

Miller, GJ and Hurley, JH. (2004) "Crystal structure of the catalytic core of inositol 1,4,5-trisphosphate 3-kinase." *Mol Cell* **15**(5):703-11.

Millward-Sadler, SJ, Davidson, K, Hazlewood, GP, Black, GW, Gilbert, HJ and Clarke, JH. (1995) "Novel cellulose-binding domains, NodB homologues and conserved modular architecture in xylanases from the aerobic soil bacteria *Pseudomonas fluorescens* subsp. *cellulosa* and *Cellvibrio mixtus*." *Biochem J* **312 (Pt 1)**:39-48.

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

Miziorko, HM. (2000) "Phosphoribulokinase: current perspectives on the structure/function basis for regulation and catalysis." *Adv Enzymol Relat Areas Mol Biol* **74**:95-127.

Mok, KH and Han, KH. (1999) "NMR solution conformation of an antitoxic analogue of α-conotoxin GI: identification of a common nicotinic acetylcholine receptor α1-subunit binding surface for small ligands and α-conotoxins." *Biochemistry* **38**(37):11895-904.

Morais, MC, Zhang, W, Baker, AS, Zhang, G, Dunaway-Mariano, D and Allen, KN. (2000) "The crystal structure of *Bacillus cereus* phosphonoacetaldehyde hydrolase: insight into catalysis of phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily." *Biochemistry* **39**(34):10385-96.

Morera, S, Lascu, I, Dumas, C, LeBras, G, Briozzo, P, Veron, M and Janin, J. (1994) "Adenosine 5'-diphosphate binding and the active site of nucleoside diphosphate kinase." *Biochemistry* **33**(2):459-67.

Morrissette, J, Kratzschmar, J, Haendler, B, el-Hayek, R, Mochca-Morales, J, Martin, BM, Patel, JR, Moss, RL, Schleuning, WD, Coronado, R, et al. (1995) "Primary structure and properties of helothermine, a peptide toxin that blocks ryanodine receptors." *Biophys J* **68**(6):2280-8.

Mueller, EJ, Oh, S, Kavalerchik, E, Kappock, TJ, Meyer, E, Li, C, Ealick, SE and Stubbe, J. (1999) "Investigation of the ATP binding site of *Escherichia coli* aminoimidazole ribonucleotide synthetase using affinity labeling and site-directed mutagenesis." *Biochemistry* **38**(31):9831-9.

Muller, YA, Heiring, C, Misselwitz, R, Welfle, K and Welfle, H. (2002) "The cystine knot promotes folding and not thermodynamic stability in vascular endothelial growth factor." *J Biol Chem* **277**(45):43410-6.

Murzin, AG. (1996) "Structural classification of proteins: new superfamilies." *Curr Opin Struct Biol* **6**(3):386-94.

Murzin, AG, Brenner, SE, Hubbard, T and Chothia, C. (1995) "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *J Mol Biol* **247**(4):536-40.

Nagano, N, Orengo, CA and Thornton, JM. (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.

Nishino, T, Komori, K, Ishino, Y and Morikawa, K. (2003) "X-ray and biochemical anatomy of an archaeal XPF/Rad1/Mus81 family nuclease: similarity between its endonuclease domain and restriction enzymes." *Structure (Camb)* **11**(4):445-57.

Nobelmann, B and Lengeler, JW. (1996) "Molecular analysis of the *gat* genes from *Escherichia coli* and of their roles in galactitol transport and metabolism." *J Bacteriol* **178**(23):6790-5.

Nobile, M, Noceti, F, Prestipino, G and Possani, LD. (1996) "Helothermine, a lizard venom toxin, inhibits calcium current in cerebellar granules." *Exp Brain Res* **110**(1):15-20.

Novotny, M, Madsen, D and Kleywegt, GJ. (2004) "Evaluation of protein fold comparison servers." *Proteins* **54**(2):260-70.

Omecinsky, DO, Holub, KE, Adams, ME and Reily, MD. (1996) "Three-dimensional structure analysis of μ-agatoxins: further evidence for common motifs among neurotoxins with diverse ion channel specificities." *Biochemistry* **35**(9):2836-44.

Orengo, CA, Michie, AD, Jones, S, Jones, DT, Swindells, MB and Thornton, JM. (1997) "CATH--a hierarchic classification of protein domain structures." *Structure* **5**(8):1093-108.

Ortiz, AR, Strauss, CE and Olmea, O. (2002) "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison." *Protein Sci* **11**(11):2606-21.

Oudot, C, Cortay, JC, Blanchet, C, Laporte, DC, Di Pietro, A, Cozzone, AJ and Jault, JM. (2001) "The "catalytic" triad of isocitrate dehydrogenase kinase/phosphatase from *E. coli* and its relationship with that found in eukaryotic protein kinases." *Biochemistry* **40**(10):3047-55.

Paakkonen, K, Tossavainen, H, Permi, P, Kilpelainen, I and Guntert, P. (*To be published*) "Structures of the First and Fourth Tsr Domains of F-Spondin."

Pandit, SB, Gosar, D, Abhiman, S, Sujatha, S, Dixit, SS, Mhatre, NS, Sowdhamini, R and Srinivasan, N. (2002) "SUPFAM--a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes." *Nucleic Acids Res* **30**(1):289-93.

Patte, JC, Clepet, C, Bally, M, Borne, F, Mejean, V and Foglino, M. (1999) "ThrH, a homoserine kinase isozyme with in vivo phosphoserine phosphatase activity in *Pseudomonas aeruginosa*." *Microbiology* **145**(4):845-53.

Paz Moreno-Murciano, M, Monleon, D, Marcinkiewicz, C, Calvete, JJ and Celda, B. (2003) "NMR solution structure of the non-RGD disintegrin obtustatin." *J Mol Biol* **329**(1):135-45.

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

Pei, J, Sadreyev, R and Grishin, NV. (2003) "PCMA: fast and accurate multiple sequence alignment based on profile consistency." *Bioinformatics* **19**(3):427-8.

Peisach, D, Gee, P, Kent, C and Xu, Z. (2003) "The crystal structure of choline kinase reveals a eukaryotic protein kinase fold." *Structure (Camb)* **11**(6):703-13.

Pennington, MW, Lanigan, MD, Kalman, K, Mahnir, VM, Rauer, H, McVaugh, CT, Behm, D, Donaldson, D, Chandy, KG, Kem, WR, et al. (1999) "Role of disulfide bonds in the structure and potassium channel blocking activity of ShK toxin." *Biochemistry* **38**(44):14549-58.

Perona, JJ, Tsu, CA, Craik, CS and Fletterick, RJ. (1993) "Crystal structures of rat anionic trypsin complexed with the protein inhibitors APPI and BPTI." *J Mol Biol* **230**(3):919-33.

Ponyi, T, Szabo, L, Nagy, T, Orosz, L, Simpson, PJ, Williamson, MP and Gilbert, HJ. (2000) "Trp22, Trp24, and Tyr8 play a pivotal role in the binding of the family 10 cellulose-binding module from *Pseudomonas* xylanase A to insoluble ligands." *Biochemistry* **39**(5):985-91.

Raghothama, S, Eberhardt, RY, Simpson, P, Wigelsworth, D, White, P, Hazlewood, GP, Nagy, T, Gilbert, HJ and Williamson, MP. (2001) "Characterization of a cellulosome dockerin domain from the anaerobic fungus *Piromyces equi*." *Nat Struct Biol* **8**(9):775-8.

Raghothama, S, Simpson, PJ, Szabo, L, Nagy, T, Gilbert, HJ and Williamson, MP. (2000) "Solution structure of the CBM10 cellulose binding module from *Pseudomonas* xylanase A." *Biochemistry* **39**(5):978-84.

Rajashankar, KR, Kniewel, R, Solorzano, V and Lima, CD. (*To be published*) "Glycerate Kinase from *Neisseria meningitidis* (Serogroup A)."

Ramon-Maiques, S, Marina, A, Uriarte, M, Fita, I and Rubio, V. (2000) "The 1.5 Å resolution crystal structure of the carbamate kinase-like carbamoyl phosphate synthetase from the hyperthermophilic Archaeon *Pyrococcus furiosus*, bound to ADP, confirms that this thermostable enzyme is a carbamate kinase, and provides insight into substrate binding and stability in carbamate kinases." *J Mol Biol* **299**(2):463-76.

Rao, VD, Misra, S, Boronenkov, IV, Anderson, RA and Hurley, JH. (1998) "Structure of type IIβ phosphatidylinositol phosphate kinase: a protein kinase fold flattened for interfacial phosphorylation." *Cell* **94**(6):829-39.

Rawlings, ND, Tolle, DP and Barrett, AJ. (2004) "MEROPS: the peptidase database." *Nucleic Acids Res* **32**(Database issue):D160-4.

Reily, MD, Holub, KE, Gray, WR, Norris, TM and Adams, ME. (1994) "Structure-activity relationships for P-type calcium channel-selective ω-agatoxins." *Nat Struct Biol* **1**(12):853-6.

Reizer, J, Ramseier, TM, Reizer, A, Charbit, A and Saier, MH, Jr. (1996) "Novel phosphotransferase genes revealed by bacterial genome sequencing: a gene cluster encoding a putative N-acetylgalactosamine metabolic pathway in *Escherichia coli*." *Microbiology* **142**:231-50.

Robinson, VL, Buckler, DR and Stock, AM. (2000) "A tale of two components: a novel kinase and a regulatory switch." *Nat Struct Biol* **7**(8):626-33.

Rossmann, MG, Moras, D and Olsen, KW. (1974) "Chemical and biological evolution of nucleotide-binding protein." *Nature* **250**(463):194-9.

Ruepp, A, Zollner, A, Maier, D, Albermann, K, Hani, J, Mokrejs, M, Tetko, I, Guldener, U, Mannhaupt, G, Munsterkotter, M, et al. (2004) "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." *Nucleic Acids Res* **32**(18):5539-45.

Russell, RB and Barton, GJ. (1994) "Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility." *J Mol Biol* **244**(3):332-50.

Russell, RB, Saqi, MA, Sayle, RA, Bates, PA and Sternberg, MJ. (1997) "Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation." *J Mol Biol* **269**(3):423-39.

Ryazanov, AG, Ward, MD, Mendola, CE, Pavur, KS, Dorovkov, MV, Wiedmann, M, Erdjument-Bromage, H, Tempst, P, Parmer, TG, Prostko, CR, et al. (1997) "Identification of a new class of protein kinases represented by eukaryotic elongation factor-2 kinase." *Proc Natl Acad Sci U S A* **94**(10):4884-9.

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

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

Sanchez, JF, Hoh, F, Strub, MP, Aumelas, A and Dumas, C. (2002) "Structure of the cathelicidin motif of protegrin-3 precursor: structural insights into the activation mechanism of an antimicrobial protein." *Structure (Camb)* **10**(10):1363-70.

Santos, MA, Jimenez, A and Revuelta, JL. (2000) "Molecular characterization of FMN1, the structural gene for the monofunctional flavokinase of *Saccharomyces cerevisiae*." *J Biol Chem* **275**(37):28618-24.

Saraste, M, Sibbald, PR and Wittinghofer, A. (1990) "The P-loop--a common motif in ATP- and GTP-binding proteins." *Trends Biochem Sci* **15**(11):430-4.

Sawano, Y, Muramatsu, T, Hatano, K, Nagata, K and Tanokura, M. (2002) "Characterization of genomic sequence coding for bromelain inhibitors in pineapple and expression of its recombinant isoform." *J Biol Chem* **277**(31):28222-7.

Schenk, B, Fernandez, F and Waechter, CJ. (2001) "The ins(ide) and out(side) of dolichyl phosphate biosynthesis and recycling in the endoplasmic reticulum." *Glycobiology* **11**(5):61R-70R.

Schlichting, I and Reinstein, J. (1999) "pH influences fluoride coordination number of the AlFx phosphoryl transfer transition state analog." *Nat Struct Biol* **6**(8):721-3.

Schultz, J, Milpetz, F, Bork, P and Ponting, CP. (1998) "SMART, a simple modular architecture research tool: identification of signaling domains." *Proc Natl Acad Sci U S A* **95**(11):5857-64.

Scordis, P, Flower, DR and Attwood, TK. (1999) "FingerPRINTScan: intelligent searching of the PRINTS motif database." *Bioinformatics* **15**(10):799-806.

Senda, T, Yamada, T, Sakurai, N, Kubota, M, Nishizaki, T, Masai, E, Fukuda, M and Mitsuidagger, Y. (2000) "Crystal structure of NADH-dependent ferredoxin reductase component in biphenyl dioxygenase." *J Mol Biol* **304**(3):397-410.

Sheng, XR, Li, X and Pan, XM. (1999) "An iso-random Bi Bi mechanism for adenylate kinase." *J Biol Chem* **274**(32):22238-42.

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

Shin, J, Hong, SY, Chung, K, Kang, I, Jang, Y, Kim, DS and Lee, W. (2003) "Solution structure of a novel disintegrin, salmosin, from *Agkistrondon halys* venom." *Biochemistry* **42**(49):14408-15.

Shirakihara, Y and Evans, PR. (1988) "Crystal structure of the complex of phosphofructokinase from *Escherichia coli* with its reaction products." *J Mol Biol* **204**(4):973-94.

Siebold, C, Arnold, I, Garcia-Alles, LF, Baumann, U and Erni, B. (2003) "Crystal structure of the *Citrobacter freundii* dihydroxyacetone kinase reveals an eight-stranded α-helical barrel ATP-binding domain." *J Biol Chem* **278**(48):48236-44.

Sigrell, JA, Cameron, AD, Jones, TA and Mowbray, SL. (1998) "Structure of *Escherichia coli* ribokinase in complex with ribose and dinucleotide determined to 1.8 Å resolution: insights into a new family of kinase structures." *Structure* **6**(2):183-93.

Sigrist, CJ, Cerutti, L, Hulo, N, Gattiker, A, Falquet, L, Pagni, M, Bairoch, A and Bucher, P. (2002) "PROSITE: a documented database using patterns and profiles as motif descriptors." *Brief Bioinform* **3**(3):265-74.

Singh, SK, Yang, K, Karthikeyan, S, Huynh, T, Zhang, X, Phillips, MA and Zhang, H. (2004) "The thrH gene product of *Pseudomonas aeruginosa* is a dual activity enzyme with a novel phosphoserine:homoserine phosphotransferase activity." *J Biol Chem* **279**(13):13166-73.

Smit, A and Mushegian, A. (2000) "Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway." *Genome Res* **10**(10):1468-84.

Song, J, Gilquin, B, Jamin, N, Drakopoulou, E, Guenneugues, M, Dauplais, M, Vita, C and Menez, A. (1997) "NMR solution structure of a two-disulfide derivative of charybdotoxin: structural evidence for conservation of scorpion toxin α/β motif and its hydrophobic side chain packing." *Biochemistry* **36**(13):3760-6.

Sonnhammer, EL, Eddy, SR and Durbin, R. (1997) "Pfam: a comprehensive database of protein domain families based on seed alignments." *Proteins* **1997**(28):3.

Sonnhammer, EL and Kahn, D. (1994) "Modular arrangement of proteins as inferred from analysis of homology." *Protein Sci* **3**(3):482-92.

Spronk, AM, Yoshida, H and Wood, HG. (1976) "Isolation of 3-phosphohistidine from phosphorylated pyruvate, phosphate dikinase." *Proc Natl Acad Sci U S A* **73**(12):4415-9.

St Charles, R, Padmanabhan, K, Arni, RV, Padmanabhan, KP and Tulinsky, A. (2000) "Structure of tick anticoagulant peptide at 1.6 Å resolution complexed with bovine pancreatic trypsin inhibitor." *Protein Sci* **9**(2):265-72.

Stammers, DK, Achari, A, Somers, DO, Bryant, PK, Rosemond, J, Scott, DL and Champness, JN. (1999) "2.0 Å X-ray structure of the ternary complex of 7,8-dihydro-6-hydroxymethylpterinpyrophosphokinase from *Escherichia coli* with ATP and a substrate analogue." *FEBS Letters* **456**(1):49-53.

Steinbacher, S, Hof, P, Eichinger, L, Schleicher, M, Gettemans, J, Vandekerckhove, J, Huber, R and Benz, J. (1999) "The crystal structure of the *Physarum polycephalum* actin-fragmin kinase: an atypical protein kinase with a specialized substrate-binding domain." *EMBO J* **18**(11):2923-9.

Stuckey, JA and Dixon, JE. (1999) "Crystal structure of a phospholipase D family member." *Nat Struct Biol* **6**(3):278-84.

Suresh, S, Turley, S, Opperdoes, FR, Michels, PA and Hol, WG. (2000) "A potential target enzyme for trypanocidal drugs revealed by the crystal structure of NAD-dependent glycerol-3-phosphate dehydrogenase from *Leishmania mexicana*." *Structure Fold Des* **8**(5):541-52.

Tan, K, Duquette, M, Liu, JH, Dong, Y, Zhang, R, Joachimiak, A, Lawler, J and Wang, JH. (2002) "Crystal structure of the TSP-1 type 1 repeats: a novel layered fold and its biological implication." *J Cell Biol* **159**(2):373-82.

Tanaka, T, Saha, SK, Tomomori, C, Ishima, R, Liu, D, Tong, KI, Park, H, Dutta, R, Qin, L, Swindells, MB, et al. (1998) "NMR structure of the histidine kinase domain of the *E. coli* osmosensor EnvZ." *Nature* **396**(6706):88-92.

Tao, Y, Farsetta, DL, Nibert, ML and Harrison, SC. (2002) "RNA synthesis in a cage--structural studies of reovirus polymerase λ3." *Cell* **111**(5):733-45.

Tari, LW, Matte, A, Goldie, H and Delbaere, LT. (1997) "Mg^{2+}-Mn^{2+} clusters in enzyme-catalyzed phosphoryl-transfer reactions." *Nat Struct Biol* **4**(12):990-4.

Tatusov, RL, Fedorova, ND, Jackson, JD, Jacobs, AR, Kiryutin, B, Koonin, EV, Krylov, DM, Mazumder, R, Mekhedov, SL, Nikolskaya, AN, et al. (2003) "The COG database: an updated version includes eukaryotes." *BMC Bioinformatics* **4**(1):41.

Tatusov, RL, Galperin, MY, Natale, DA and Koonin, EV. (2000) "The COG database: a tool for genome-scale analysis of protein functions and evolution." *Nucleic Acids Res* **28**(1):33-6.

Tatusov, RL, Koonin, EV and Lipman, DJ. (1997) "A genomic perspective on protein families." *Science* **278**(5338):631-7.

Tatusov, RL, Natale, DA, Garkavtsev, IV, Tatusova, TA, Shankavaram, UT, Rao, BS, Kiryutin, B, Galperin, MY, Fedorova, ND and Koonin, EV. (2001) "The COG database: new developments in phylogenetic classification of proteins from complete genomes." *Nucleic Acids Res* **29**(1):22-8.

Taylor, WR and Orengo, CA. (1989) "Protein structure alignment." *J Mol Biol* **208**(1):1-22.

Thomas, PD, Kejariwal, A, Campbell, MJ, Mi, H, Diemer, K, Guo, N, Ladunga, I, Ulitsky-Lazareva, B, Muruganujan, A, Rabkin, S, et al. (2003) "PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification." *Nucleic Acids Res* **31**(1):334-41.

Thompson, TB, Thomas, MG, Escalante-Semerena, JC and Rayment, I. (1998) "Three-dimensional structure of adenosylcobinamide kinase/adenosylcobinamide phosphate guanylyltransferase from *Salmonella typhimurium* determined to 2.3 Å resolution." *Biochemistry* **37**(21):7686-95.

Thornton, JM. (1981) "Disulphide bridges in globular proteins." *J Mol Biol* **151**(2):261-87.

van de Locht, A, Stubbs, MT, Bode, W, Friedrich, T, Bollschweiler, C, Hoffken, W and Huber, R. (1996) "The ornithodorin-thrombin crystal structure, a key to the TAP enigma?" *EMBO J* **15**(22):6011-7.

van Mierlo, CP, Darby, NJ, Neuhaus, D and Creighton, TE. (1991) "(14-38, 30-51) double-disulphide intermediate in folding of bovine pancreatic trypsin inhibitor: a two-dimensional 1H nuclear magnetic resonance study." *J Mol Biol* **222**(2):353-71.

van Vlijmen, HW, Gupta, A, Narasimhan, LS and Singh, J. (2004) "A novel database of disulfide patterns and its application to the discovery of distantly related homologs." *J Mol Biol* **335**(4):1083-92.

Vaughn, JL, Feher, V, Naylor, S, Strauch, MA and Cavanagh, J. (2000) "Novel DNA binding domain and genetic regulation model of *Bacillus subtilis* transition state regulator abrB." *Nat Struct Biol* **7**(12):1139-46.

Vita, C, Drakopoulou, E, Vizzavona, J, Rochette, S, Martin, L, Menez, A, Roumestand, C, Yang, YS, Ylisastigui, L, Benjouad, A, et al. (1999) "Rational engineering of a miniprotein that reproduces the core of the CD4 site interacting with HIV-1 envelope glycoprotein." *Proc Natl Acad Sci U S A* **96**(23):13091-6.

Walker, DR and Koonin, EV. (1997) "SEALS: a system for easy analysis of lots of sequences." *Proc Conf Intell Syst Mol Biol* **5**:333-9.

Walker, JE, Saraste, M, Runswick, MJ and Gay, NJ. (1982) "Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-

requiring enzymes and a common nucleotide binding fold." *EMBO J* **1**(8):945-51.

Wallon, G, Kryger, G, Lovett, ST, Oshima, T, Ringe, D and Petsko, GA. (1997) "Crystal structures of *Escherichia coli* and *Salmonella typhimurium* 3-isopropylmalate dehydrogenase and comparison with their thermophilic counterpart from *Thermus thermophilus*." *J Mol Biol* **266**(5):1016-31.

Wang, J, Shen, B, Guo, M, Lou, X, Duan, Y, Cheng, XP, Teng, M, Niu, L, Liu, Q, Huang, Q, et al. (2005) "Blocking Effect and Crystal Structure of Natrin Toxin, a Cysteine-Rich Secretory Protein from *Naja atra* Venom that Targets the BK_{Ca} Channel." *Biochemistry* **44**(30):10145-10152.

Wang, W, Kim, R, Jancarik, J, Yokota, H and Kim, SH. (2001) "Crystal structure of phosphoserine phosphatase from *Methanococcus jannaschii*, a hyperthermophile, at 1.8 Å resolution." *Structure* **9**(1):65-71.

Watson, HC, Walker, NP, Shaw, PJ, Bryant, TN, Wendell, PL, Fothergill, LA, Perkins, RE, Conroy, SC, Dobson, MJ, Tuite, MF, et al. (1982) "Sequence and structure of yeast phosphoglycerate kinase." *EMBO J* **1**(12):1635-40.

Wierenga, RK. (2001) "The TIM-barrel fold: a versatile framework for efficient enzymes." *FEBS Letters* **492**(3):193-8.

Wilding, EI, Brown, JR, Bryant, AP, Chalker, AF, Holmes, DJ, Ingraham, KA, Iordanescu, S, So, CY, Rosenberg, M and Gwynn, MN. (2000) "Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci." *J Bacteriol* **182**(15):4319-27.

Wistow, G and Piatigorsky, J. (1987) "Recruitment of enzymes as lens structural proteins." *Science* **236**(4808):1554-6.

Wolf, YI, Grishin, NV and Koonin, EV. (2000) "Estimating the number of protein folds and families from complete genome data." *J Mol Biol* **299**(4):897-905.

Wu, CH, Nikolskaya, A, Huang, H, Yeh, LS, Natale, DA, Vinayaka, CR, Hu, ZZ, Mazumder, R, Kumar, S, Kourtesis, P, et al. (2004) "PIRSF: family classification system at the Protein Information Resource." *Nucleic Acids Res* **32**(Database issue):D112-4.

Xiang, Y, Huang, RH, Liu, XZ, Zhang, Y and Wang, DC. (2004) "Crystal structure of a novel antifungal protein distinct with five disulfide bridges from *Eucommia ulmoides* Oliver at an atomic resolution." *J Struct Biol* **148**(1):86-97.

Xu, J, Li, M, Lin, G, Kim, D and Xu, Y. (2003) "Protein threading by linear programming." *Pac Symp Biocomput*:264-75.

Xu, YW, Morera, S, Janin, J and Cherfils, J. (1997) "AlF_3 mimics the transition state of protein phosphorylation in the crystal structure of nucleoside diphosphate kinase and MgADP." *Proc Natl Acad Sci U S A* **94**(8):3579-83.

Yamaguchi, H, Matsushita, M, Nairn, AC and Kuriyan, J. (2001) "Crystal structure of the atypical protein kinase domain of a TRP channel with phosphotransferase activity." *Mol Cell* **7**(5):1047-57.

Yan, Y and Moult, J. (2005) "Protein Family Clustering for Structural Genomics." *J Mol Biol* **353**(3):744-59.

Yang, D, Shipman, LW, Roessner, CA, Scott, AI and Sacchettini, JC. (2002) "Structure of the *Methanococcus jannaschii* mevalonate kinase - a member of the GHMP kinase superfamily." *J Biol Chem* **277**(11):9462-7.

Yankovskaya, V, Horsefield, R, Tornroth, S, Luna-Chavez, C, Miyoshi, H, Leger, C, Byrne, B, Cecchini, G and Iwata, S. (2003) "Architecture of succinate dehydrogenase and reactive oxygen species generation." *Science* **299**(5607):700-4.

Yano, H, Kuroda, S and Buchanan, BB. (2002) "Disulfide proteome in the analysis of protein function and structure." *Proteomics* **2**(9):1090-6.

Yasutake, Y, Watanabe, S, Yao, M, Takada, Y, Fukunaga, N and Tanaka, I. (2002) "Structure of the monomeric isocitrate dehydrogenase: evidence of a protein monomerization by a domain duplication." *Structure (Camb)* **10**(12):1637-48.

Yun, M, Park, CG, Kim, JY, Rock, CO, Jackowski, S and Park, HW. (2000) "Structural basis for the feedback regulation of *Escherichia coli* pantothenate kinase by coenzyme A." *J Biol Chem* **275**(36):28093-9.

Zdobnov, EM and Apweiler, R. (2001) "InterProScan--an integration platform for the signature-recognition methods in InterPro." *Bioinformatics* **17**(9):847-8.

Zeth, K, Ravelli, RB, Paal, K, Cusack, S, Bukau, B and Dougan, DA. (2002) "Structural analysis of the adaptor protein ClpS in complex with the N-terminal domain of ClpA." *Nat Struct Biol* **9**(12):906-11.

Zhou, A, Huntington, JA, Pannu, NS, Carrell, RW and Read, RJ. (2003) "How vitronectin binds PAI-1 to modulate fibrinolysis and cell migration." *Nat Struct Biol* **10**(7):541-4.

Zhou, G, Somasundaram, T, Blanc, E, Parthasarathy, G, Ellington, WR and Chapman, MS. (1998) "Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions." *Proc Natl Acad Sci U S A* **95**(15):8449-54.

Zhou, T, Daugherty, M, Grishin, NV, Osterman, AL and Zhang, H. (2000) "Structure and mechanism of homoserine kinase: prototype for the GHMP kinase superfamily." *Structure* **8**(12):1247-57.

Zhu, Y, Huang, W, Lee, SS and Xu, W. (2005) "Crystal structure of a polyphosphate kinase and its implications for polyphosphate synthesis." *EMBO Rep* **6**(7):681-7.

VITAE

Sara Anne Cheek was born in Indianapolis, Indiana on July 24, 1978. She is the daughter of Michael and Janet Cheek. In 1996, she graduated as the salutatorian of her class from Center Grove High School, Greenwood, Indiana. She attended Purdue University, West Lafayette, Indiana and received the degree of Bachelor of Science with a major in biology and a minor in mathematics in May 2000. She entered the Graduate School of Biomedical Sciences at the University of Texas Southwestern Medical Center in Dallas, Texas in the fall of 2000 and joined the laboratory of Dr. Nick Grishin in the spring of 2001.

Permanent Address: 1127 Old Eagle Way
Greenwood, IN 46143