Structural Classification of Thioredoxin-Like Fold Proteins

Yuan Qi[2] and Nick V. Grishin[1, 2*]
[1]*Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas*
[2]*Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas*

ABSTRACT **Protein structure classification is necessary to comprehend the rapidly growing structural data for better understanding of protein evolution and sequence–structure–function relationships. Thioredoxins are important proteins that ubiquitously regulate cellular redox status and various other crucial functions. We define the thioredoxin-like fold using the structure consensus of thioredoxin homologs and consider all circular permutations of the fold. The search for thioredoxin-like fold proteins in the PDB database identified 723 protein domains. These domains are grouped into eleven evolutionary families based on combined sequence, structural, and functional evidence. Analysis of the protein–ligand structure complexes reveals two major active site locations for the thioredoxin-like proteins. Comparison to existing structure classifications reveals that our thioredoxin-like fold group is broader and more inclusive, unifying proteins from five SCOP folds, five CATH topologies and seven DALI domain dictionary globular folding topologies. Considering these structurally similar domains together sheds new light on the relationships between sequence, structure, function and evolution of thioredoxins. Proteins 2005;58:376–388.**

Key words: fold definition; circular permutation; homology; active site location; structural analog; structure-based multiple sequence alignment

INTRODUCTION

More than 24,000 experimentally determined protein structures have been deposited to the Protein Data Bank (PDB)[1] and the rate increase in the growth of structure data is anticipated as high-throughput structural genomics continues.[2] To comprehend this large amount of data for better understanding of protein evolution and sequence–structure–function relationships, protein structure classification is necessary. Several hierarchical protein structure classifications exist, with the major ones being SCOP[3–5], CATH[6–8] and Dali Domain Dictionary (DaliDD)[9–11]. In a protein structure classification, fold group and evolutionary family are the two major levels. At the fold level, protein domains are grouped based on the connectivity and mutual orientation of their core secondary structure elements. Within each fold group, proteins are further divided into evolutionary families based on inferred homology relationships. The geometry of protein structures usually reflects certain constraints from sequence and function. Thus grouping proteins by folds will aid in understanding of the physico-chemical principles behind protein structures, which in turn could help to address problems such as protein folding and structure–functional prediction. Furthermore, although a few exceptional examples exist where homologous proteins have evolved different folds,[12,13] protein structures generally evolve slower than their sequences. Consequently, homologous proteins could share the same fold and other subtle structural features even when their sequences have diverged beyond recognition. Therefore, grouping protein domains by folds could also help in understanding protein evolution and will facilitate homology inference.

A systematic comparison of the three major structure classifications (SCOP, CATH, DaliDD) shows many discrepancies, even at the fold-group level.[14] These discrepancies create obstacles for homology inference and modeling, evolutionary studies, and genome annotation. One major source of the inconsistencies stems from the concept of fold definition. Structural fold concept is a perception of a researcher and thus is intrinsically subjective. The definition of a protein fold is therefore somewhat arbitrary. For example, it is difficult to define and to distinguish folds of regular-layered architectures, especially α/β sandwiches. Their β-sheets take up a large proportion of the structure and are similar due to hydrogen-bonding constraints, and the differences between structures could be only a few secondary structure elements.[15,16] In an effort to understand and to clarify fold definitions for proteins with α/β sandwich architectures, we start from a large and diverse protein group, namely thioredoxin-like proteins.

Thioredoxin is an important redox protein that is present in every organism. Together with thioredoxin reductase and peroxiredoxin, thioredoxin regulates the cellular reduction/oxidation status as well as various important cellular functions, such as oxidative stress defense, cell proliferation, signal transduction, and transcription regulation.[17–21] Extensive studies have been done on thioredoxin.[17–22] Consequently, a large number of X-ray and NMR struc-

Grant sponsor: The National Institutes of Health; Grant number: GM67165.

*Correspondence to: Nick V. Grishin, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390-9050. E-mail: grishin@chop.swmed.edu

Received 1 March 2004; Accepted 29 April 2004

Published online 19 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20329

tures are available for thioredoxin and related proteins, rendering their classification necessary.

Figure 1(c) shows the structure of a human thioredoxin, which is a three-layer α/β/α sandwich with the central β-sheet formed by five β-strands flanked by two α-helices on each side. Many proteins important for cellular thiol-redox pathways, such as glutaredoxin, protein disulfide isomerase (PDI) and oxidase (DsbA), and glutathione S-transferase (GST), are homologous to thioredoxin and have similar structures. However, many of these classical thioredoxin-like proteins do not contain α-helix α0′ and β-strand β0′, and some do not contain α-helix α3′ [Fig. 1(c)]. To generate a consistent and inclusive definition of the thioredoxin-like fold, we use the structure consensus of thioredoxins and the classical thioredoxin-like proteins that are undoubtedly homologs to each other, and only include those secondary-structure elements and interactions that are present in all these homologs [Fig. 1(a)]. Interestingly, a circularly permuted DsbA protein exists as a result of a protein engineering experiment that is structurally stable and functionally active.[23] As homologous proteins can evolve to have different circular permutations (e.g., DNA methyltransferases[24]), we decide not to limit our fold-group definition to identical topology, but to consider all potential circular permutations of the thioredoxin-like fold.

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

MATERIALS AND METHODS

Structural Motif Search for Thioredoxin-Like Protein Domains

We used a structure motif search program (unpublished) that was under development in our lab. Briefly, the program generated a database of PDB structures (19,558 structures, July 2003), in which each structure was represented by a Secondary Structure Element Interaction Matrix describing the interactions (parallel or anti-parallel), hydrogen-bonding and chirality between the secondary structure elements of the PDB structure. The structure consensus of the classical thioredoxin-like proteins (thioredoxin, glutaredoxin, protein disulfide isomerases (PDI), disulfide bond oxidase (DsbA), glutathione S-transferase (GST), glutathione peroxidase, and their close homologs)[25] was represented as a query matrix. The query matrix [Fig. 1(b)] specified the number and types of secondary structure elements in the thioredoxin motif, the hydrogen-bonding and parallel or anti-parallel relationships between the four β-strands, and the chirality between consecutive secondary structures. We then used our structure-motif search program to search the database of Secondary Structure Element Interaction Matrices of every PDB structure and to output the structures containing submatrix matching the query matrix. Six query matrices characterizing six possible circular permutations of the thioredoxin motif were constructed and searched for. False positives were removed by visual inspection. Proteins were considered to contain the thioredoxin-motif only when the thioredoxin motif formed the structural core of the protein domain (see "Structural analogs" section for details).

Classification of the Thioredoxin-Like Protein Domains

The thioredoxin motif-containing protein domains retrieved as described above were subsequently grouped into evolutionary families using a combined sequence, structural and functional analysis. We used four methods to search for sequence similarities between the thioredoxin motif-containing domains and all PDB proteins: gapped BLAST,[26,27] PSI-BLAST,[27,28] RPS-BLAST,[29] and COMPASS,[30] each of which uses a query sequence or profile to search a database of sequences or profiles. A query sequence was the sequence of every thioredoxin-motif containing domain. A query profile was generated by running a query sequence against the nr database (1,479,768 sequences, 476,959,297 total letters, Aug 2003) using PSI-BLAST for up to five iterations with an inclusion E-value cutoff of 0.005. The database of PDB sequences contained sequences of PDB chains (49,319 sequences, 10,645,968 total letters, Aug 2003). The database of domain profiles contained the profiles of representative protein domains in the PDB. We used the SCOP v1.63 domain definitions for this purpose. The representative SCOP v1.63 domain sequences with less than 40% sequence identity to each other (5,224 domains) were downloaded from Astral.[31,32] A profile for each representative domain sequence was then generated in the same way as we generated a query profile. We searched each query sequence in the database of PDB sequences using Gapped BLAST,[26,27] each query profile in the database of PDB sequences using PSI-BLAST,[27,28] each query sequence in the database of domain profiles using RPS-BLAST,[29] and each query profile in the database of domain profiles using COMPASS.[30] Sequence analyses were based on the search results of the four methods. We also inspected each hit with an E-value up to 10 so that we would not miss a potential homolog that has a signature sequence motif but with a less significant E-value.

For structure analysis, 723 thioredoxin motif-containing protein domains were first clustered according to their sequence identities using the program BLASTCLUST (I. Dondoshansky and Y. Wolf, unpublished; ftp://ftp.ncbi.nih.gov/blast/) at a sequence identity threshold of 50% and length coverage of 90%. A representative structure for each cluster was selected based on the quality of the structure (resolution, R factor value, solved date for NMR structures) and the presence of ligands or substrate analogs. All structure analyses were done on this set of the

representative domain structures. The representative structures were aligned in an all-against-all manner using the program DaliLite and were further clustered by a Dali Z-score cutoff of five. The representative structures were visualized in the INSIGHT II package (MSI) and superimposed by aligning structurally equivalent residues. A structure-based multiple sequence alignment of all 90 representative structures was constructed manually taking into account alignments made by DaliLite,[33] Mammoth,[34] CE,[35] PSI-BLAST[27,28] and RPS-BLAST.[29] The structural alignment was further filtered by sequence identities in the aligned regions and the final alignment contained proteins that had less than 50% sequence identity to each other. The ligands or substrate analogs and active site residues were also visualized in INSIGHT II and locations of active sites were compared.

RESULTS AND DISCUSSION

Overall Fold Description

Thioredoxin-like fold

Many proteins important for cellular thiol-redox pathways, such as thioredoxin, glutaredoxin, glutathione S-transferase (GST), protein disulfide bond isomerase (PDI), are known to adopt the thioredoxin-like fold.[25] In both SCOP and CATH, the thioredoxin/glutaredoxin fold is described as a three-layer α/β/α sandwich. As shown in Figure 1(c), thioredoxin is a three-layer sandwich with a central β-sheet flanked by two α-helices on each side. However, the N-terminal α-helix, α0′ is absent in many classical thioredoxin-like fold proteins, such as GST, bacterial glutaredoxin, and archaeon PDI. α-Helix α3′ is also not conserved in the thioredoxin homologs. For instance, the N-terminal domain of a bacterial alkyl hydroperoxide reductase subunit F (1hyuA1), which is a close homolog of PDI, has only a short loop connecting β2 and β3 in the place of the α-helix α3′ (Fig. 2). In addition, phosducin (1a0rP), a homolog of thioredoxin, has only a loop with turns in the place of the α-helix α3′ (Fig. 2). In many proteins that do have α-helices at the α3′ position, these α-helices are irregular, kinked or appear as separated short helical turns. Based on these observations, the first α-layer of the thioredoxin fold is not conserved in all thioredoxin homologs. Since the fold definition should include only the core secondary structural elements that are present in the majority of homologs, we define the thioredoxin-like fold as a two-layer α/β sandwich with the βαββββα secondary-structure pattern. The four β-strands, ordering 2134, form a mixed β-sheet with the third β-strand anti-parallel to the rest, and the two α-helices packed against the β-sheet on one side [Fig. 1(a)]. The N-terminal half of the fold is a right-handed βαβ unit. This unit is connected through a loop to the C-terminal half of the fold, which is a β-hairpin followed by an α-helix and the chirality of this ββα unit is left-handed. Consequently, the chiralities between secondary structure elements β4, α2, β1, and α2, β1, α1 are both right-handed.

Applying this definition, we searched for all potential thioredoxin-like protein domains in the entire PDB database using the structure motif search program under development in our lab. Found proteins containing the βαββββα unit with the thioredoxin-like interactions (see Materials and Methods and Fig. 1) were visually inspected to ensure that the six elements form the structural core (see "Structural analogs" section for clarification) of the protein domains. Altogether 723 protein domains were identified as thioredoxin-like fold proteins. They were unified into the thioredoxin-like fold group and divided into evolutionary families. A structure-based multiple sequence alignment of 90 representative thioredoxin-like fold protein domains was manually constructed (Fig. 2). From this alignment, we see that some thioredoxin-like proteins have insertions of secondary structure elements into the common structural motif. A number of proteins from four families possess the α-helix α3′. Proteins from other four families have an extra αβ unit inserted between the β-strands β2 and β3, extending the central β-sheet to be formed by five β-strands.

Circular permutations

The protein domains that we unified into the thioredoxin-like fold group represent different circular permutations of the thioredoxin-like motif. A circular permutation of a structural motif can be visualized as an imaginary "ligation" of the N- and C- termini followed by an imaginary "cleavage" at a loop region of the motif to create different termini. Except when specifically mentioned, we use the phrase "circular permutation" only to indicate this kind of geometric relationship between structures and not to imply evolutionary events. It has been documented, however, that circular permutations occur in nature as evolutionary scenarios and represent a mechanism of potential fold change in evolution.[24,36–38] Since proteins with different circular permutations of a structural motif have essentially the same spatial arrangement of secondary structure elements, the same side-chain packing interactions,

Fig. 1. Thioredoxin-like fold and its observed circular permutations. **a:** The topological diagram of the thioredoxin-like fold. α-helices and β-strands are shown as blue cylinders and yellow arrows, respectively, the lines connecting different secondary structures represent loop regions between them. Dotted loops indicate the termini positions of the four types of circular permutations that we observed. No termini were observed at solid loop locations. Loops shown in red indicate the type i active site location. **b:** The query matrix of the thioredoxin-like fold of type I circular permutation. Secondary structures are consecutively numbered in Arabic numbers. Upper case letters E (β-strand) and H (α-helix) indicate the type of secondary structure. Lower case letters c and t indicate parallel and anti-parallel hydrogen-bonding interactions between secondary structures, respectively. Upper case letter X indicates that no interactions were considered. Upper case letters R and L indicate right-handed and left-handed chirality in a triplet of secondary structures, respectively. Ribbon diagrams of (**c**) human thioredoxin (1ert[56]), a representative of type I circular permutation, (**d**) yeast ribosomal protein L30 (1cn8A[42]), a representative of type II circular permutation, (**e**) *E. coli* cytidine deaminase (1aln_1[51]), a representative of type III circular permutation, and (**f**) bacteriophage HK97 capsid protein gp5 (1ohg[57], previous PDB ID: 1fh6), a representative of type IV circular permutation, were produced using the program MOLSCRIPT.[58] Corresponding secondary structure elements are colored and named as in diagram (a). Elements corresponding to inserted domains are shown in white. The long insertion in capsid protein gp5 is shown in purple in (f). In (c), (e), and (f), active site residues are depicted in red ball-and-stick representation. In (d), active site residues interacting with RNA are shown in red. The orange sphere in (e) shows a zinc ion.

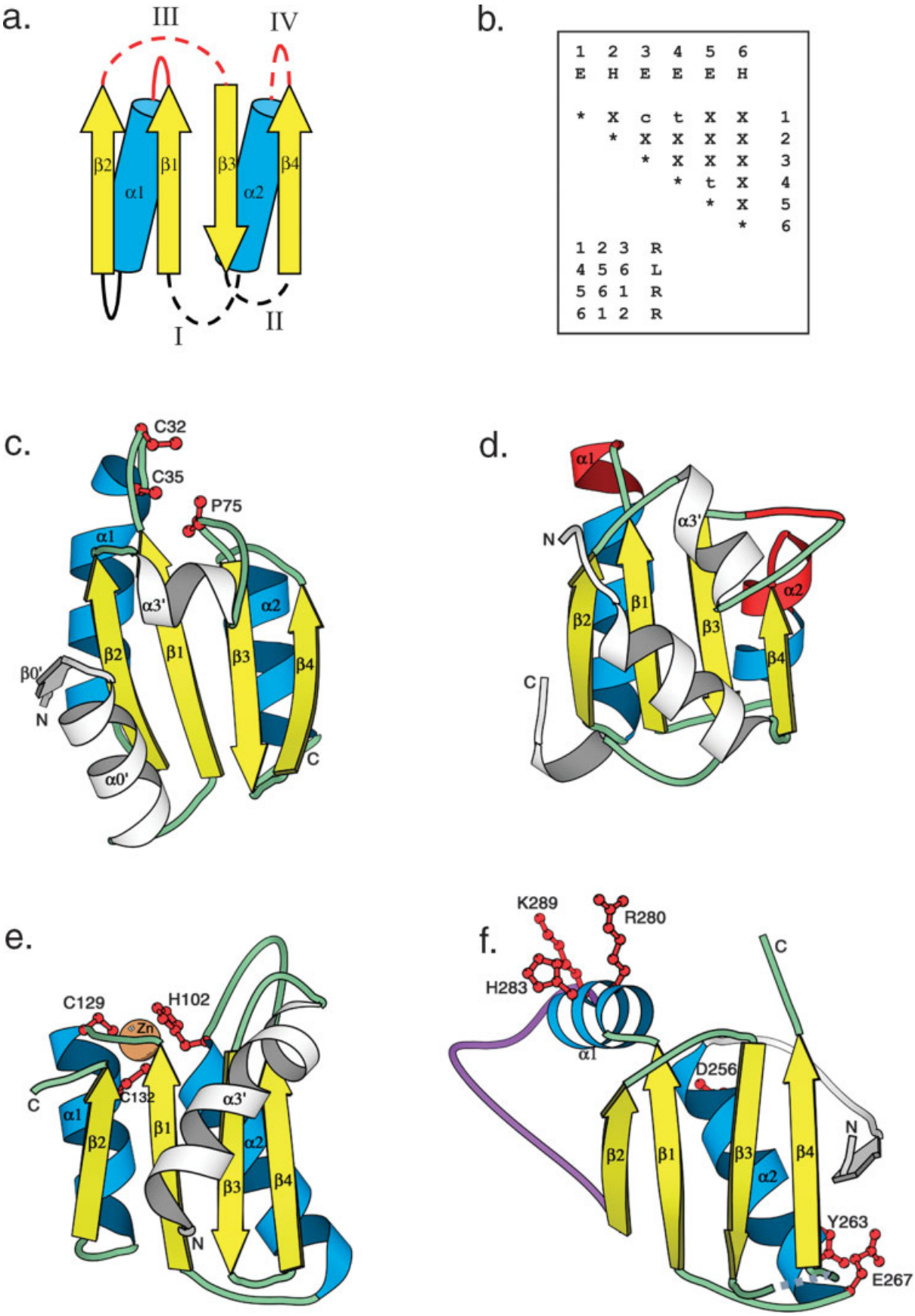
a.
III
IV
β2
α1
β1
β3
α2
β4
I
II
b.
1 2 3 4 5 6
E H E E E H
* X c t X X 1
* X X X X 2
* X X X 3
* t X 4
* X 5
* 6
1 2 3 R
4 5 6 L
5 6 1 R
6 1 2 R
c.
C32
C35
P75
α1
α3'
α2
β2
β1
β3
β4
β0'
N
C
α0'
d.
α1
α3'
N
β2
β1
β3
α2
β4
C
e.
C129
H102
Zn
C132
C
α1
β2
β1
α3'
α2
β3
β4
N
f.
K289
R280
C
H283
α1
D256
β2
β1
β3
α2
β4
N
Y263
E267

Figure 1.

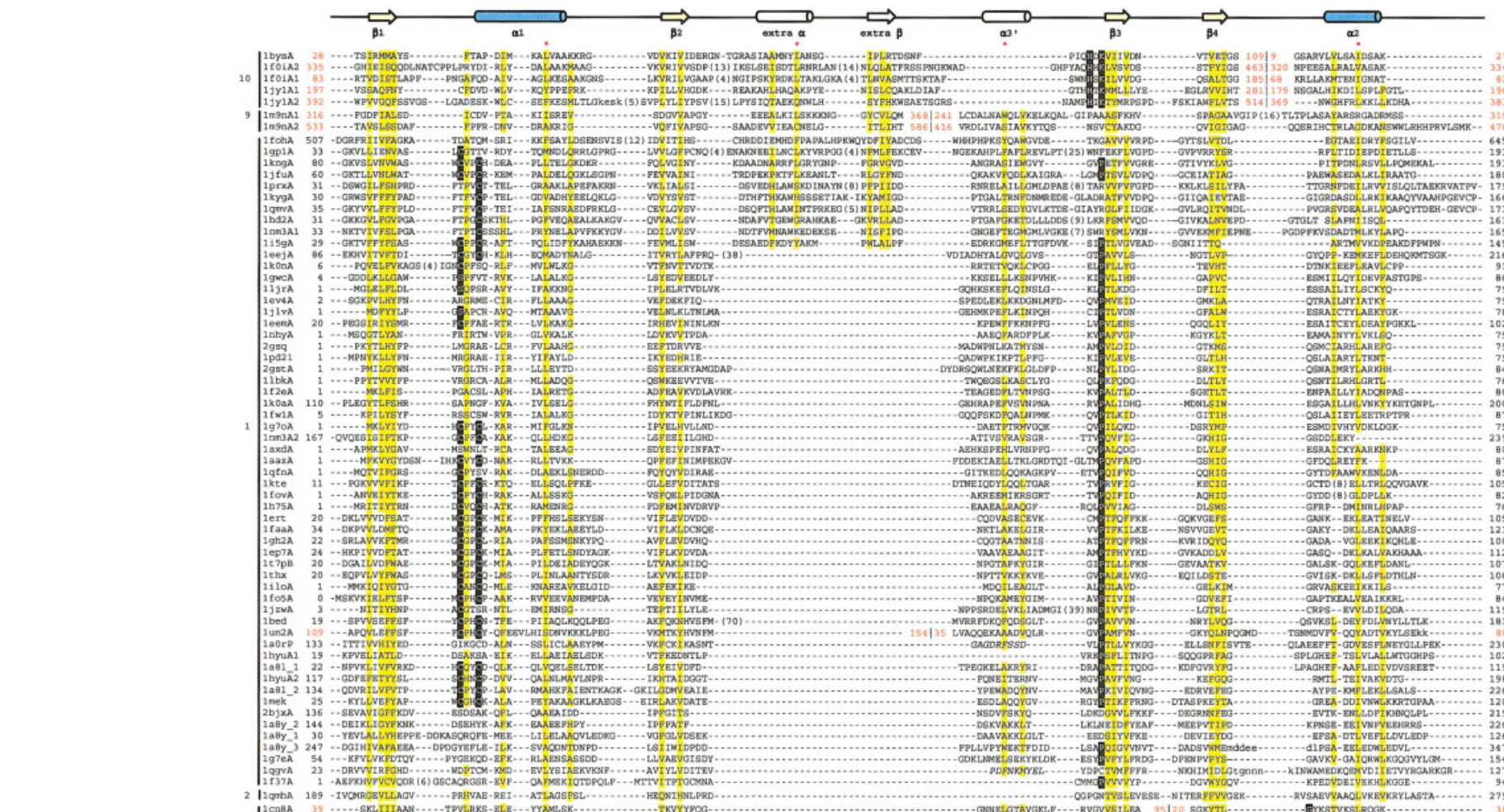

Figure 2.

and may be homologous, grouping them together into the same fold group for further comparative analysis could help us to better understand protein folding and sequence–structure–function relationships and potential evolutionary connections. We can use the structure-based multiple sequence alignment to study the sequence similarities between proteins with different circular permutations. Such potential similarities are obscured if the proteins are classified in different fold groups or even different structural classes.

Since the thioredoxin-like motif contains six secondary structure elements, six types of circular permutations are theoretically possible by placing the termini before each secondary structure element. However, only four types of circular permutations were seen in the PDB database [Fig. 1(a)]. No proteins are present with the termini positioned between β1–α1 or α1–β2, suggesting that β1α1β2 may be an essential folding or packing unit for the thioredoxin-like fold. This observation agrees with the finding by Salem et al. that βαβ-unit is one of the three most prominent (highly-populated) supersecondary structures.[39] However, it is possible that with more structures accumulating in the PDB, circular permutation variants that disrupt the βαβ-unit will appear. Out of the four types of circular permutations we see, type II (β4α2β1α1β2β3; secondary structures are numbered the same as those in the classical thioredoxin-like proteins) is adopted in five families and the other three types are all adopted in two families, respectively (Table I). If we count the number of representative structures, type I (β1α1β2β3β4α2) is the most populated, and type II is the second-most populated type of circular permutation.

Description of Thioredoxin-Like Fold Families

We identified 723 protein domains as belonging to the thioredoxin-like fold. We subsequently classified these protein domains into eleven evolutionary families based on inferred homology relationships between them. While we gathered strong support for homology of protein domains within each evolutionary family, we are not drawing any conclusion about the evolutionary relationship between protein domains in different families. Protein domains from different families could simply be analogous to each other. Alternatively, they could share homologous relationships that we were not able to support convincingly, or be mosaics of homologous and analogous pieces. It has been hypothesized that modern protein domains have evolved from combinations of ancient domain segments composed of supersecondary structures, and thioredoxin fold proteins is a possible example of such domain evolution.[40] Although a detailed analysis of this problem is very challenging and lies beyond the scope of our current study, this evolutionary scenario is plausible. However, we believe that for the proteins within each of our evolutionary families, homologous segments spans through the entire common core of the domain. Here we describe the eleven families and discuss their sequence, structural and functional features with evolutionary implications. The representatives of each family are listed in Figure 2.

Thioredoxin family

This family includes all the classical thioredoxin-like proteins as well as calsequestrin, phosducin, and arsenate reductase, among others. The dithiol-disulfide oxidoreductases, such as thioltransferases and PDI, have a conserved active-site sequence motif Cys-X-X-Cys that is located at the N-terminus of α-helix α1. In addition, a cis-proline residue located at the loop region before β3 is conserved and is in spatial proximity to the Cys-X-X-Cys motif [Fig. 1(c)]. Proteins that form inter-domain disulfide bonds, such as glutathione peroxidases, and proteins that do not form disulfide bonds, such as the N-terminal domain of

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

TABLE I.

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

[a]Previous PDB identifier: 1dyv
[b]Previous PDB identifier: 1fh6

elongation factor 1-gamma (eEF1gamma), have lost one or both of the conserved Cys residues (Fig. 2). Nevertheless, they have the same active site locations as the dithiol-disulfide oxidoreductases, and their homology relationships with the dithiol-disulfide oxidoreductases can be inferred from PSI-BLAST and RPS-BLAST results and close structural similarities.

Protein domains in this family have a type I circular permutation except for one disulfide bond oxidase (DsbA, 1un2A, previous PDB ID: 1dyv) that is a type III circular permutation as the result of a protein engineering experiment.[23] Aside from the common structural motif, most thioredoxins and PDIs have the extra α-helices α0′ and α3′ [Fig. 1(c)]. Glutathione peroxidases and peroxiredoxins have an extra α/β unit inserted between β2 and β3 and the extra β-strand is hydrogen-bonded with β2 (Fig. 2); DsbAs have an extra β-strand inserted before β1 and hydrogen-bonded with β4; so they all have a mixed β-sheet of five β-strands.

RTPC small domain family

Similar to thioredoxins, the small domains of the RNA 3′-terminal phosphate cyclases (RTPC) have the type I circular permutation. However, the β-sheet in this family is much flatter and the β-strands are up to four residues longer than those of the thioredoxins. The functional role of the RTPC small domain remains unknown.[41]

Ribosomal protein L30e family

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

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

Protein domains in the L30e family interact with their ligands and substrates at the N-terminal ends of the α-helices and nearby regions. The yeast ribosomal protein L30 interacts with the RNA internal loop through the residues located at the N-terminal ends of

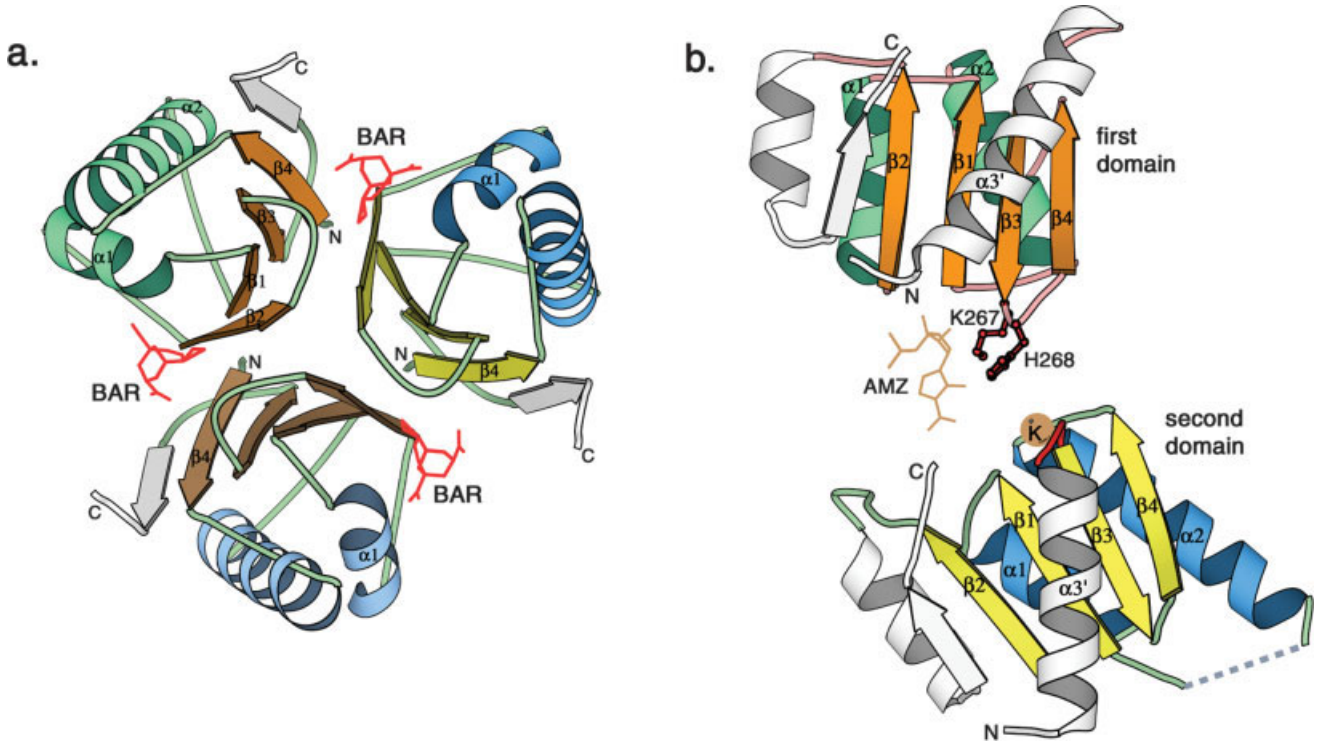

Fig. 3. Active-site locations. **a:** Ribbon diagram of bacillus chorismate mutase (2cht[45]) with its substrate analogs shows the type ii active site location. The α-helices and β-strands are numbered as in Figure 1(a), but are colored differently in different domains with inserted elements in white, and substrate analog BAR in red. The three domains are viewed along their three-fold axis. One BAR molecule is located at each of the three clefts between two adjacent domains. **b:** Ribbon diagram of two thioredoxin-like domains in the AICAR Tfase part of bifunctional purine biosynthesis enzyme ATIC (1m9n[55]) illustrates an unusual active site location. The second domain of one monomer is colored as in Figure 1(a), while the first domain of another monomer is shown in a different color scheme. The other two thioredoxin-like domains of the AICAR Tfase part are omitted for clarity. The substrate AMZ is shown in brown and marks the active site. Two catalytic residues from the first domain are shown as ball-and-stick in red. A potassium ion represented by an orange sphere binds to the loop (shown in red) between α3′ and β4 of the second domain. Both ribbon diagrams were generated using the program MOLSCRIPT.[58]

α-helices α1 and α2 and in the loop region before β-strand β3[42] [Fig. 1(d)].

Tubulin C-terminal domain family

This family includes the C-terminal domains of tubulin α- and β-subunit, cell division protein FtsZ, and dihydroxyacetone kinase subunit K (DhaK). The overall structures of tubulin, FtsZ, and DhaK are similar; all are formed of two domains that have the same relative positions. In all proteins of this family, the N-terminal domains are Rossmann-like nucleotide-binding domains: GTPase for tubulin and FtsZ, and ATPase for DhaK. The C-terminal domains are the thioredoxin-like domains with a type II circular permutation. The C-terminal domain of DhaK has a β-hairpin inserted between β4 and α2. The substrate Dha is covalently bound[43] to this β-hairpin. In tubulin, the loop between β4 and α2 [Fig. 1(a)] also forms a functional site where the ligands, zinc ion, and anticancer drug taxol, bind.[44]

Bacillus chorismate mutase (BCM) family

Bacillus and *Thermus* chorismate mutase, hypothetical protein YjgF, and purine regulatory protein YabJ are placed in this family. Simple BLAST results show that *Bacillus* with *Thermus* chorismate mutases and YjgF with YabJ form two clusters of close homologs. Despite the low sequence identity (average 8.6%) between the two groups, their tertiary and quaternary structures are very similar to each other. These proteins are homotrimers; each monomer is a thioredoxin-like domain of type II circular permutation. The three β-sheets from three monomers form a barrel-shaped interface. When looking parallel to the three-fold axis that goes in the direction from the C-terminus to the N-terminus of α-helices α1 and α2, the three β-sheets of proteins in both groups run approximately parallel to the axis with a left-handed twist [Fig. 3(a)]. The monomers of *Thermus* chorismate mutase (1odeA) and YjgF (1qu9A) are superimposed with a RMSD of 1.7 Å based on 84 C_α atoms, and quaternary structures are superimposed very well. The active site locations are also the same for the two group of proteins, which are at the three clefts between adjacent monomers[45,46] [Fig. 3(a)], indicating homology.

MECP synthase family

The quaternary structures of 2C-methyl-D-erythritol-2,4-cyclodiphosphate (MECP) synthases are similar to proteins in the *Bacillus* chorismate mutase (BCM) family. MECP synthases are also homotrimers with each monomer a thioredoxin-like domain of type II circular permuta-

tion. However, there are several structural and functional site differences between MECP synthases and BCM family proteins. The monomers of MECP synthases have an extra α-helix between α2 and β1 that is absent in the BCM family proteins. The β-sheets of MECP synthases also run approximately parallel to the three-fold axis but with a right-handed twist instead of a left-handed one, so the monomer cannot superimpose well when the trimers are superimposed with the BCM family members. The active sites of MECP synthases are also located at the clefts between adjacent monomers.[47,48] However, in MECP synthases, the active site residues are contributed from β2 and α1 of one monomer and β4 and α2 of the adjacent monomer; while in BCM family proteins, the active site residues are contributed from β2 and α1 of one monomer but β3 and β4 of the adjacent monomer. These differences between MECP synthases and BCM family proteins indicate that they may not share a common ancestor. Therefore, we place MECP synthases in a separate family.

PurM N-terminal domain family

The N-terminal domain of the aminoimidazole ribonucleotide synthetase (PurM) is a thioredoxin-like domain of type II circular permutation, with an extra nine-residue α-helix inserted between β4 and α2. PurMs are homodimers, and the two β-sheets from the two thioredoxin-like N-terminal domains form a barrel-shaped dimer interface. The active site of PurM is proposed to be formed by the edge β-strands of the two β-sheets and the C-terminal domain.[49]

Cytidine deaminase family

This family includes single domain cytidine deaminase (CDA), two-domain CDA, and cytosine deaminase. The N-terminal domain of the two-domain CDA has a higher sequence identity (29%) to the single domain CDA than to its C-terminal domain (15%), suggesting that the two-domain CDA emerged by an ancient gene duplication event of the one-domain CDA and the C-terminal domain diverged further. In fact, the N-terminal domain of the two-domain CDA, the single domain CDA, and the cytosine deaminase all have two conserved cysteines at the N-terminus of α-helix α1 and a conserved cysteine or histidine at the N-terminus of α-helix α2 that coordinate a catalytic zinc ion[50,51] [Figs. 1(e), 2], while the C-terminal domain of two-domain CDA has lost the zinc coordination and thus the catalytic activity.

All domains in this family have a type III circular permutation [β3β4α2β1α1β2, Fig. 1(e)], the same as the engineered DsbA. The loop regions before β3 and between β4-α2 are about eight residues longer than most domains of the thioredoxin family (Fig. 2), and they form a cover of the hydrophobic active site. Like glutathione peroxidases, cytosine deaminase has an extra α/β unit inserted after β2 and the extra β-strand is hydrogen-bonded with β2 (Fig. 2). One-domain CDA and the N-terminal domain of two-domain CDA have an extra β-strand inserted after β2 and is also hydrogen-bonded with β2, but it is oppositely oriented compared to the extra β-strand in cytosine deaminase.

AICAR Tfase domain of bifunctional purine biosynthesis enzyme ATIC family

The bifunctional purine biosynthesis enzyme ATIC has two functional parts: the inosine monophosphate cyclohydrolase (IMPCH) part and the 5-aminoimidazole-4-carboxamide-ribonucleotide transfermylase (AICAR Tfase) part. ATIC is a homodimer with each monomer participating in both functional parts.[52] Each AICAR Tfase part of the monomer includes two thioredoxin-like domains that are structurally very similar to each other (RMSD of 1.17 Å based on 118 atoms). The two thioredoxin-like domains in the same polypeptide chain are the result of an ancient gene-duplication event, and thus they are homologous to each other. The two thioredoxin-like domains are of type III circular permutation, and like glutathione peroxidases of the thioredoxin family, each of them have an extra αβ unit inserted after β2 [Figs. 3(b), 2]. The second thioredoxin-like domain has an insertion of a small helical domain between α2 and β1. AICAR Tfase has two active sites; each is located between the first thioredoxin-like domain of one monomer and the second thioredoxin-like domain of the other monomer. Our analysis shows that the two homologous thioredoxin-like domains possess different active site locations [Fig. 3(b) and "Analysis of active-site locations," below].

Phospholipase D family

This family includes phospholipase D, bacterial nuclease Nuc, and tyrosyl-DNA phosphodiesterase (TDP1). Phospholipase D and Nuc are inferred as close homologs based on RPS- and PSI-BLAST results (RPS-BLAST E-value: 5e-14), while TDP1 was previously shown by Interthal etc.[53] to be homologous to phospholipase D and Nuc based on the presence of the conserved HK motif (Fig. 2) and similar reaction mechanism. Both phospholipase D and TDP1 contain two duplicated thioredoxin-like domains of type IV circular permutation (α2β1α1β2β3β4). Nuc only contains one such domain, but it is a homodimer and the two monomers are arranged in the same way as the two domains in phospholipase D and TDP1. Like glutathione peroxidases of the thioredoxin family, all protein domains in this family have an extra αβ unit inserted after β2 (Fig. 2).

gp5 domain A family

Domain A of the major capsid protein gp5 is a thioredoxin-like domain of type IV circular permutation. Protein gp5 is the assembly subunit of the double-strand DNA bacteriophage HK97 capsid.[54] Each capsid asymmetric unit is a hexamer or a pentamer of gp5. Other domains of gp5s, domain P, E-loop, and N-arm form a hexagon or a pentagon, and domains A of gp5s form a cover of the space inside the polygon. A 22-residue-long insertion between α-helix α1 and β-strand β2 pushes the C-terminus of α1 up and makes α1 almost perpendicular to the β-sheet instead of being parallel to it [Fig. 1(f)]. This arrangement renders α1 anti-parallel to α-helix α2 of the neighboring gp5 domain A. α1 and α2 of adjacent domains A form electro-

static and hydrophobic interactions in between to stabilize the cover of the polygon.

Analysis of Active-Site Locations

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

In many proteins, active site (type i location) is placed at the N-terminal ends of the α-helices or nearby loop regions, i.e., the binding or catalytic residues are located on the loops connecting β1-α1, β2-β3, β4-α2, or at the N-termini of the α-helices α1/α2 [Fig. 1(a)]. This type of active-site location is adopted by protein domains in five different families that encompass all four circular permutations (Table I). Since protein domains with this active site location belong to different evolutionary families, the similarity in the active site placement may be the result of convergent evolution and is probably caused by physicochemical constraints such as the helix dipoles of α1 and α2.

Another common placement of the active site (type ii location) is along the edges of the β- sheet, i.e., the binding or catalytic residues are located on the edge β-strands (β2, β4) of the β-sheet or on the sides of α-helices α1 and α2 that are facing opposite from each other. This type of active-site location is adopted by protein domains in four different families (Table I). Proteins in three of the families (*Bacillus* chorismate mutase, MECP synthase, and PurM N-terminal domain) form homo-trimers or dimers and the β-sheets of the trimer or dimer form a barrel-shaped interface. Their active sites are placed in the clefts between adjacent monomers [Fig. 3(a)] and thus are constrained to the edges of the α/β sandwich for each monomer. Although protein domains of the gp5 domain A family do not bind substrates or ligands, they do interact with each other, participating in formation of homo-hexamers or pentamers stabilized partially by electrostatic and hydrophobic interactions between α-helices α1 and α2 of adjacent monomers [Fig. 1(f)].

While homologous protein domains usually have similar active-site locations, we found an exception to this rule. As we mentioned in the family description above, the four thioredoxin-like domains of the AICAR Tfase part of the bifunctional purine biosynthesis enzyme ATIC are homologous to each other. The active site of AICAR Tfase is between the first thioredoxin-like domain of one monomer and the second thioredoxin-like domain of the other monomer [Fig. 3(b)]. The second thioredoxin-like domain houses active-site residues at the loop regions near the N-terminal ends of the α-helices, similarly to most other thioredoxin-like domains (type i location); while the first thioredoxin-like domain has active-site residues in the loop regions near the C-terminal ends of the α-helices, with two catalytic residues located at the loop between β-strands 3 and 4,[55] which is opposite to that of the second domain [Fig. 3(b)]. Thus, our analysis reveals a rare example of homologous protein domains possessing different active-site locations.

Comparison to Other Structure Classifications

Different structure classifications use different criteria and methods. The protein domains that we unified in the thioredoxin-like fold group are categorized differently in three major structure classifications CATH[6–8], SCOP[3–5], and Dali Domain Dictionary.[9–11]

In CATH (version 2.5)[6–8], some of these thioredoxin-like fold protein domains are not classified at all, such as the 2C-methyl-D-erythritol-2,4-cyclodiphosphate synthase, the AICAR transfermylase domain of bifunctional purine biosynthesis enzyme ATIC, and the capsid gp5 protein domain A; the others are placed into five fold groups (CATH "topology" level). Three of the fold groups correspond to three different circular permutations, and protein domains of type I circular permutation are divided into two fold groups. CATH assigns the small domain of RNA 3′-terminal phosphate cyclase (RTPC) to a different fold group than the thioredoxin proteins, although they both have the same type of circular permutation. In fact, CATH classifies them into two different architecture types (a higher level in the classification hierarchy than fold groups): a two-layer sandwich and a three-layer sandwich. The other fold groups are also categorized as two- or three-layer sandwich architecture types. CATH groups our thioredoxin-like protein domains into nine homologous superfamilies, which is basically consistent with our evolutionary family classification except for one protein. We assigned the C-terminal domain of phenol hydroxylase (1fohA) to be in the same evolutionary family as the classical thioredoxin-like proteins, while CATH assigns it into a separate superfamily by itself.

SCOP (version 1.65)[3–5] classifies the thioredoxin-like fold domains into five different fold groups (SCOP "fold" level) corresponding to four different circular permutations and one separate fold group for the entire capsid protein gp5. SCOP does not break gp5 into domains; instead, it assigns the entire gp5 protein to a separate fold group and describes it as an unusual fold. The small domain of RTPC is assigned to the same fold group as the thioredoxin proteins in SCOP. SCOP fold groups are placed into two different structural classes: α/β and α+β. At the evolutionary family level, our classification is consistent with SCOP superfamily classification.

Dali Domain Dictionary (DaliDD, version 3.1 beta)[9–11] classifies the thioredoxin-like fold protein domains into seven fold groups (DaliDD "globular folding topology" level). In this classification, there are protein domains of

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

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

Structural Analogs

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

Peptide methionine sulfoxide reductase (PMSR) contains two overlapping structural motifs: the thioredoxin-like motif and the ferredoxin-like motif. Figure 4(c) shows a typical ferredoxin-like fold protein. It is an α/β sandwich with the βαββαβ secondary-structure pattern. The four β-strands ordering 2314 form an anti-parallel β-sheet with the two α-helices on one side. From Figure 4(a), we can see that if we treat α-helix αA and β-strand βB as insertions, PMSR adopts a thioredoxin-like fold of type III circular permutation. On the other hand, if we treat α-helix α1 and β-strand β2 as insertions, the protein adopts a ferredoxin-like fold [Fig. 4(b)]. α1, β2 and αA, βB are placed on different sides of the central β-sheet and thus occupy similar positions in relation to the structure core. β2 and βB have approximately the same length [Fig. 4(a, b)]. Thus if we try to base our decision about the fold solely on the structural properties of this molecule, both structural motifs (thioredoxin-like and ferredoxin-like) appear reasonable and we are unable to choose one of them. Sequence analysis, however, shows that PMSR is homologous to the ferredoxin-like fold protein that is shown in Figure 4(c), the fourth metal-binding domain of Menkes copper-transporting ATPase (L.N. Kinch and N.V. Grishin, unpublished), that is missing α-helix α1 and β-strand β2 and hence missing the thioredoxin-like motif. Therefore, the ferredoxin-like motif is the essential one for PMSR, and PMSR most likely obtained the thioredoxin-like motif later in the process of evolution by insertions of α-helix α1 and β-strand β2. In spite of the thioredoxin-like motif in PMSR structure, it should be classified in a ferredoxin-like fold. Using similar reasoning we ruled out the following proteins with the thioredoxin motif: C-terminal domain of glyceraldehyde-3-phosphate dehydrogenase (1a7kA), transcription factor sc-mtTFB (1i4wA), and histidyl-tRNA synthetase (1adjA).

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

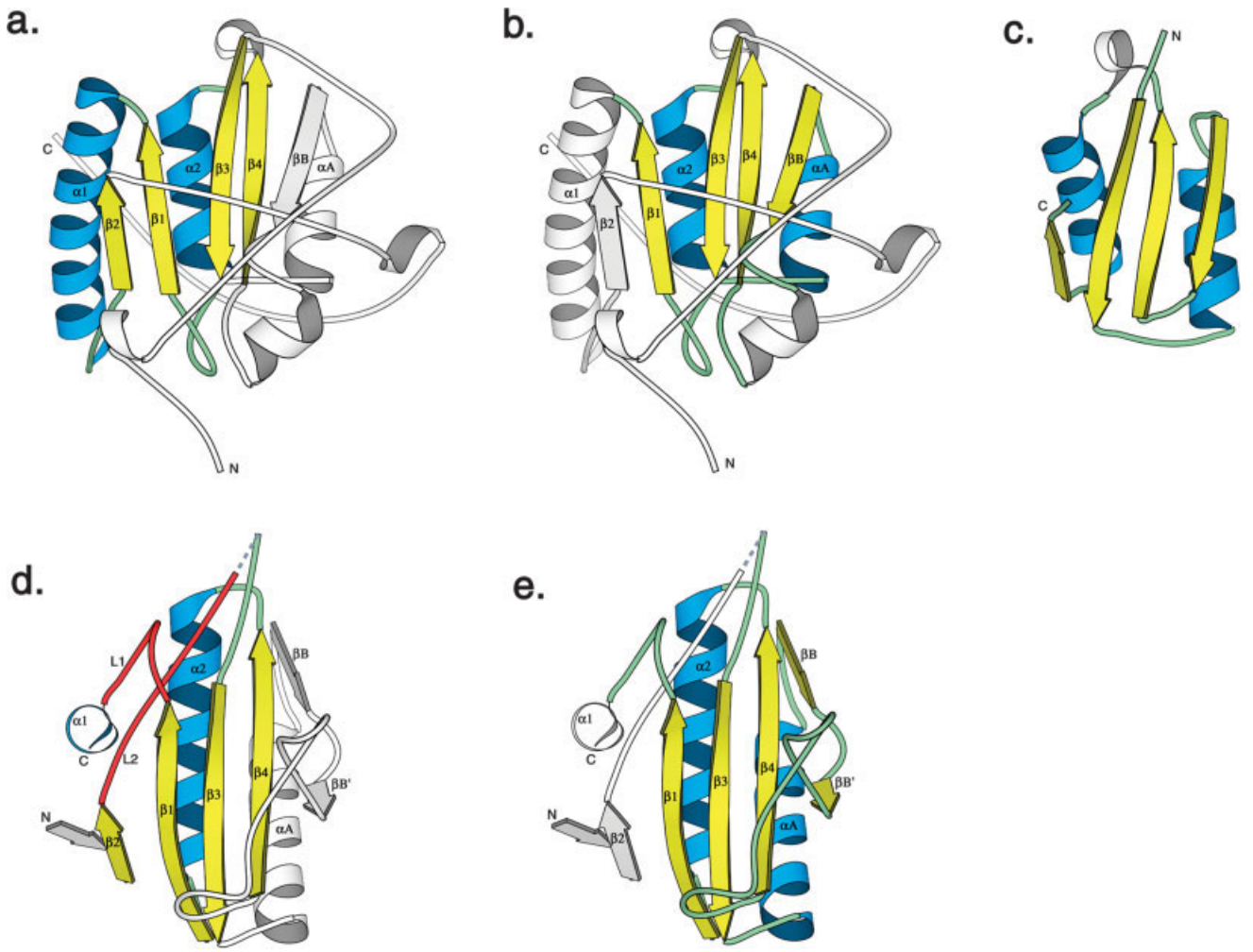

Fig. 4. Structure analogs. Ribbon diagrams of (**a**, **b**) *E. coli* peptide methionine sulfoxide reductase (1ff3[59]), (**c**) the fourth metal-binding domain of human Menkes copper-transporting ATPase (1aw0[60]), a ferredoxin-like fold protein, and (**d**, **e**) archaeon formylmethanofuran:tetrahydromethanopterin formyltransferase (1ftr[61]). Protein domains in (a) and (b) are the same, however, in (a), the elements of the thioredoxin-like motif are colored in yellow and blue; in (b), the elements of the ferredoxin-like motif are colored in yellow and blue. Similarly, we colored the 1ftr domain in (d) and (e). In (d), the loops L1 and L2 are shown in red. All ribbon diagrams were generated using the program MOLSCRIPT.[58]

ACKNOWLEDGMENTS

We thank Alexander Pertsemlidis, Lisa Kinch, and James Wrabl for critical reading of the manuscript and helpful comments. This work was supported by NIH grant GM67165 to NVG.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The Protein Data Bank and structural genomics. Nucleic Acids Res 2003;31: 489–491.
3. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
4. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
5. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
6. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
7. Pearl FM, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA. Assigning genomic sequences to CATH. Nucleic Acids Res 2000;28:277–282.
8. Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D, Pearl FM, Sillitoe I, Todd AE, Thornton JM. The CATH protein family database: a resource for structural and functional annotation of genomes. Proteomics 2002;2:11–21.
9. Dietmann S, Holm L. Identification of homology in protein structure classification. Nat Struct Biol 2001;8:953–957.
10. Holm L, Sander C. Dictionary of recurrent domains in protein structures. Proteins 1998;33:88–96.
11. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
12. Grishin NV. KH domain: one motif, two folds. Nucleic Acids Res 2001;29:638–643.
13. Murzin AG. How far divergent evolution goes in proteins. Curr Opin Struct Biol 1998;8:380–387.
14. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure Fold Des 1999;7:1099–112.
15. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification

and classification of protein fold families. Protein Eng 1993;6:485–500.
16. Orengo CA, Sillitoe I, Reeves G, Pearl FM. Review: what can structural classifications reveal about protein evolution? J Struct Biol 2001;134:145–165.
17. Nakamura H, Nakamura K, Yodoi J. Redox regulation of cellular activation. Annu Rev Immunol 1997;15:351–369.
18. Arner ES, Holmgren A. Physiological functions of thioredoxin and thioredoxin reductase. Eur J Biochem 2000;267:6102–6109.
19. Yamawaki H, Haendeler J, Berk BC. Thioredoxin: a key regulator of cardiovascular homeostasis. Circ Res 2003;93:1029–1033.
20. Das KC. Thioredoxin system in premature and newborn biology. Antioxid Redox Signal 2004;6:177–184.
21. Kontou M, Will RD, Adelfalk C, Wittig R, Poustka A, Hirsch-Kauffmann M, Schweiger M. Thioredoxin, a regulator of gene expression. Oncogene 2004;23:2146–2152.
22. Holmgren A. Thioredoxin structure and mechanism: conformational changes on oxidation of the active-site sulfhydryls to a disulfide. Structure 1995;3:239–243.
23. Hennecke J, Sebbel P, Glockshuber R. Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. J Mol Biol 1999;286:1197–215.
24. Jeltsch A. Circular permutations in the molecular evolution of DNA methyltransferases. J Mol Evol 1999;49:161–164.
25. Martin JL. Thioredoxin—a fold for all reasons. Structure 1995;3:245–250.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
28. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
29. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 2002;30:281–283.
30. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003;326:317–336.
31. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28:254–256.
32. Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. ASTRAL compendium enhancements. Nucleic Acids Res 2002;30:260–263.
33. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16:566–567.
34. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
35. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
36. Ponting CP, Russell RB. Swaposins: circular permutations within genes encoding saposin homologues. Trends Biochem Sci 1995;20:179–180.
37. Gong W, O'Gara M, Blumenthal RM, Cheng X. Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. Nucleic Acids Res 1997;25:2702–2715.
38. Bujnicki JM. Sequence permutations in the molecular evolution of DNA methyltransferases. BMC Evol Biol 2002;2:3.
39. Salem GM, Hutchinson EG, Orengo CA, Thornton JM. Correlation of observed fold frequency with the occurrence of local structural motifs1. Journal of Molecular Biology 1999;287:969–981.
40. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 2001;134:191–203.
41. Palm GJ, Billy E, Filipowicz W, Wlodawer A. Crystal structure of RNA 3′-terminal phosphate cyclase, a ubiquitous enzyme with unusual topology. Structure Fold Des 2000;8:13–23.
42. Mao H, White SA, Williamson JR. A novel loop-loop recognition motif in the yeast ribosomal protein L30 autoregulatory RNA complex. Nat Struct Biol 1999;6:1139–1147.
43. Siebold C, Garcia-Alles LF, Erni B, Baumann U. A mechanism of covalent substrate binding in the x-ray structure of subunit K of the Escherichia coli dihydroxyacetone kinase. Proc Natl Acad Sci USA 2003;100:8188–8192.
44. Lowe J, Li H, Downing KH, Nogales E. Refined structure of alpha beta-tubulin at 3.5 Å resolution. J Mol Biol 2001;313:1045–1057.
45. Chook YM, Ke H, Lipscomb WN. Crystal structures of the monofunctional chorismate mutase from Bacillus subtilis and its complex with a transition state analog. Proc Natl Acad Sci USA 1993;90:8600–8603.
46. Chook YM, Gray JV, Ke H, Lipscomb WN. The monofunctional chorismate mutase from *Bacillus subtilis*. Structure determination of chorismate mutase and its complexes with a transition state analog and prephenate, and implications for the mechanism of the enzymatic reaction. J Mol Biol 1994;240:476–500.
47. Kemp LE, Bond CS, Hunter WN. Structure of 2C-methyl-D-erythritol 2,4- cyclodiphosphate synthase: an essential enzyme for isoprenoid biosynthesis and target for antimicrobial drug development. Proc Natl Acad Sci USA 2002;99:6591–6596.
48. Kishida H, Wada T, Unzai S, Kuzuyama T, Takagi M, Terada T, Shirouzu M, Yokoyama S, Tame JR, Park SY. Structure and catalytic mechanism of 2-C-methyl-D-erythritol 2,4-cyclodiphosphate (MECDP) synthase, an enzyme in the non-mevalonate pathway of isoprenoid synthesis. Acta Crystallogr D Biol Crystallogr 2003;59:23–31.
49. Li C, Kappock TJ, Stubbe J, Weaver TM, Ealick SE. X-ray crystal structure of aminoimidazole ribonucleotide synthetase (PurM), from the Escherichia coli purine biosynthetic pathway at 2.5 Å resolution. Structure Fold Des 1999;7:1155–1166.
50. Johansson E, Mejlhede N, Neuhard J, Larsen S. Crystal structure of the tetrameric cytidine deaminase from Bacillus subtilis at 2.0 Å resolution. Biochemistry 2002;41:2563–2570.
51. Xiang S, Short SA, Wolfenden R, Carter CW, Jr. Cytidine deaminase complexed to 3-deazacytidine: a "valence buffer" in zinc enzyme catalysis. Biochemistry 1996;35:1335–1341.
52. Greasley SE, Horton P, Ramcharan J, Beardsley GP, Benkovic SJ, Wilson IA. Crystal structure of a bifunctional transformylase and cyclohydrolase enzyme in purine biosynthesis. Nat Struct Biol 2001;8:402–406.
53. Interthal H, Pouliot JJ, Champoux JJ. The tyrosyl-DNA phosphodiesterase Tdp1 is a member of the phospholipase D superfamily. Proc Natl Acad Sci USA 2001;98:12009–12014.
54. Wikoff WR, Liljas L, Duda RL, Tsuruta H, Hendrix RW, Johnson JE. Topologically linked protein rings in the bacteriophage HK97 capsid. Science 2000;289:2129–2133.
55. Wolan DW, Greasley SE, Beardsley GP, Wilson IA. Structural insights into the avian AICAR transformylase mechanism. Biochemistry 2002;41:15505–15513.
56. Weichsel A, Gasdaska JR, Powis G, Montfort WR. Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. Structure 1996;4:735–751.
57. Helgstrand C, Wikoff WR, Duda RL, Hendrix RW, Johnson JE, Liljas L. The refined structure of a protein catenane: the HK97 bacteriophage capsid at 3.44 Å resolution. J Mol Biol 2003;334:885–899.
58. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.
59. Tete-Favier F, Cobessi D, Boschi-Muller S, Azza S, Branlant G, Aubry A. Crystal structure of the Escherichia coli peptide methionine sulphoxide reductase at 1.9 Å resolution. Structure Fold Des 2000;8:1167–1178.
60. Gitschier J, Moffat B, Reilly D, Wood WI, Fairbrother WJ. Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATPase. Nat Struct Biol 1998;5:47–54.
61. Ermler U, Merckel M, Thauer R, Shima S. Formylmethanofuran: tetrahydromethanopterin formyltransferase from Methanopyrus kandleri—new insights into salt-dependence and thermostability. Structure 1997;5:635–646.