PROTEINS: Structure, Function, and Genetics 52:323–331 (2003)

PREDICTION REPORT

Deciphering a Novel Thioredoxin-Like Fold Family

Lisa N. Kinch,[1] David Baker,[2] and Nick V. Grishin[1]*
[1]*Howard Hughes Medical Institute, and Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas*
[2]*Department of Biochemistry, University of Washington, Seattle, Washington*

ABSTRACT **Sequence- and structure-based searching strategies have proven useful in the identification of remote homologs and have facilitated both structural and functional predictions of many uncharacterized protein families. We implement these strategies to predict the structure of and to classify a previously uncharacterized cluster of orthologs (COG3019) in the thioredoxin-like fold superfamily. The results of each searching method indicate that thioltransferases are the closest structural family to COG3019. We substantiate this conclusion using the ab initio structure prediction method rosetta, which generates a thioredoxin-like fold similar to that of the glutaredoxin-like thioltransferase (NrdH) for a COG3019 target sequence. This structural model contains the thiol-redox functional motif CYS-X-X-CYS in close proximity to other absolutely conserved COG3019 residues, defining a novel thioredoxin-like active site that potentially binds metal ions. Finally, the rosetta-derived model structure assists us in assembling a global multiple-sequence alignment of COG3019 with two other thioredoxin-like fold families, the thioltransferases and the bacterial arsenate reductases (ArsC). Proteins 2003;52:323–331.** © 2003 Wiley-Liss, Inc.*

Key words: thioltransferase; thiol redox; homology detection; PSI-BLAST; fold recognition; structure prediction; rosetta; threading; protein classification

INTRODUCTION

Prediction of the structures and functions of new or uncharacterized protein families is often based on remote homology to known proteins. To achieve such predictions, sensitive profile-based sequence similarity searches such as PSI-BLAST[1] and HMMer[2] are useful tools. However, when sequence diversity within a protein family is low, sequence similarity between protein families is low, or many insertions or deletions exist between protein families, these sequence-based methods provide only marginal statistics to support homology. In such cases, we can infer homology transitively by finding an "intermediate" sequence or sequence group that links to both protein families.[3,4] Alternatively, structure-based searching or fold-recognition methods that consider evolutionary relatedness[5–7] can further extend detection limits, because of the prevalence of substantial structural similarities in the absence of significant sequence identities between many protein families. Just as the success of sequence-based searching strategies depends heavily on the composition of the sequence database, the success of fold-recognition or threading methods depends on the composition of the structure database.

Having the potential to produce protein structures that do not exist in current databases, ab initio protein structure prediction can potentially overcome the limitations of traditional fold-recognition methods.[8,9] As assessed by CASP4, rosetta outperforms other ab initio methods on several novel fold targets and in some cases surpasses traditional fold-recognition algorithms on fold targets with architectures similar to known structures.[8,9] Rosetta assumes that similar sequence segments found in different proteins will adopt similar local structural conformations.[10,11] In the procedure, Monte Carlo simulation is used to construct a large number of independent fold decoys from a library of small structural fragments (3- or 9-residue segments) that display sequence similarity to a target sequence profile.[12,13] The resulting structures are clustered based on root-mean-square deviation (RMSD), and the centers of the largest clusters represent the highest confidence models.[14] In this article, we employ rosetta to create a structural model of a target sequence belonging to an orthologous group of proteins (COG3019) with unknown structure and function. We use this model

Abbreviations: amino acids are abbreviated with standard three-letter codes; ArsC, arsenate reductase; CASP4, critical assessment of protein structure prediction; COG, clusters of orthologous groups; DsbA, disulfide reductase; HMMer, hidden Markov model; PDI and DsbC, protein disulfide isomerases; PSI-BLAST, position-specific iterated basic local alignment search tool

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

Received 28 August 2002; Accepted 23 January 2003

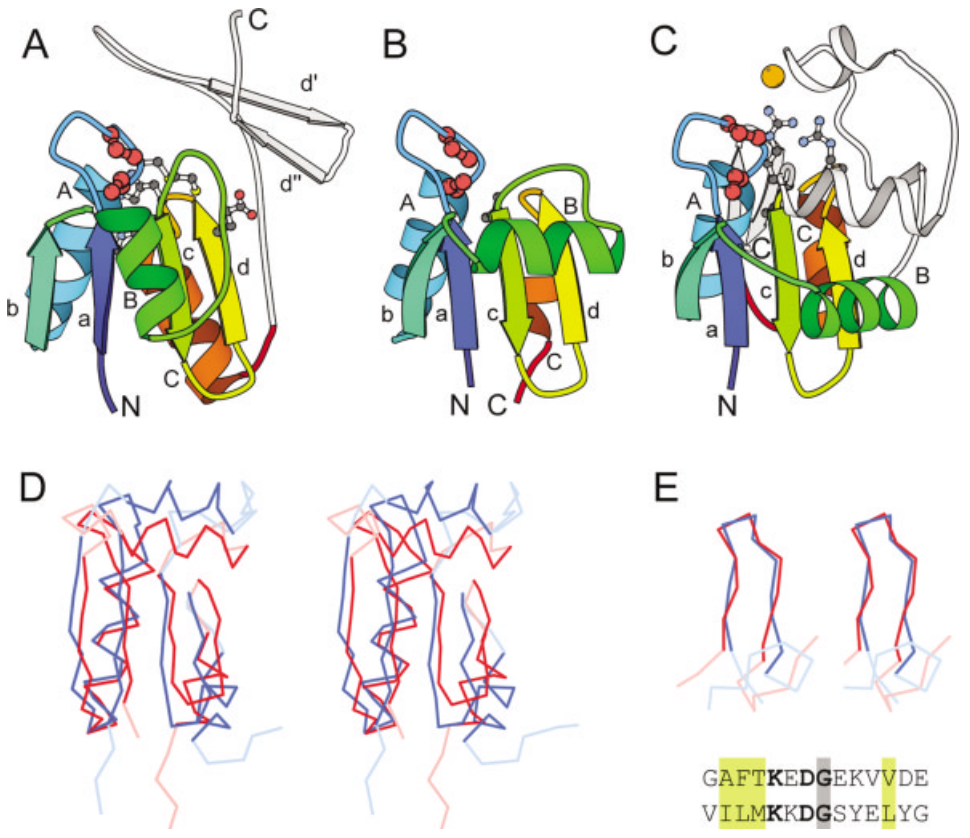

Figure 1.

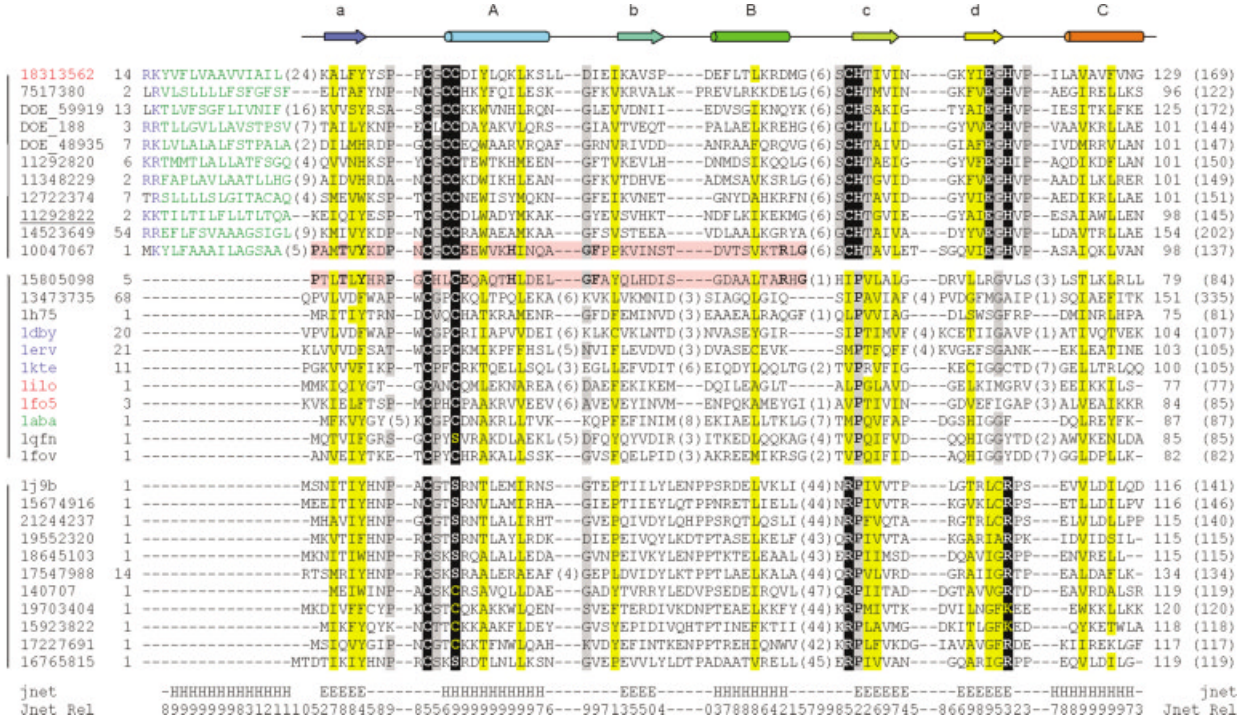

Figure 2.

both to support the remote homology-based prediction that COG3019 belongs to the thioredoxin-like fold superfamily and to generate a structure-based global multiple-sequence alignment of the families.

The structural classification of proteins (SCOP[15]) describes the thioredoxin fold as a three-layer αβα sandwich. The fold contains a mixed, four-stranded, mainly parallel β-sheet flanked by two helices on one side and a third helix on the other (Fig. 1). Members of the thioredoxin-like fold superfamily share a common core βαβαββα secondary structural pattern, with different insertions of secondary structural elements or domains distinguishing the various structural families. Many of the thioredoxin-like families contain members that function in cellular thiol-redox pathways by maintaining the reduction or the oxidation of protein or small-molecule disulfide bonds. Such members include the thioltransferases thioredoxin and glutaredoxin, the disulfide bond isomerases (PDI and DsbC) and oxidases (DsbA), the glutathione S-transferases, the glutathione peroxidase-like proteins, and the bacterial arsenate reductase (ArsC).

In addition to their common architecture, the thiol-disulfide oxidoreductases retain an active-site sequence motif CYS-X-X-CYS, with the first Cys residue of the motif distinguishing thiol-redox function. The motif occupies the loop preceding the first α-helix of the core thioredoxin-like fold (helix A, Fig. 1). A conserved proline residue resides in close structural proximity to this motif (the N-terminus of C, Fig. 1). This proline residue adopts a *cis* conformation and plays an important role in both the structure and the function of thiol oxidoreductases.[16] Other members of the of thioredoxin-like fold superfamily, such as phosducin, calsequestrin, and thioredoxin-like 2Fe-2S ferredoxin, have lost these active-site residues and do not perform thiol-redox reactions.

By combining sensitive sequence comparison methods with fold-recognition and ab initio structure prediction methods, we identify COG3019 as a new member of the thioredoxin-like fold superfamily. Using this information, we infer homology between protein sequences of this group and those belonging to the thioltransferase family of the thioredoxin-like fold superfamily. The presence of a signal peptide leader sequence and of a conserved CYS-X-X-CYS motif in COG3019 sequences suggests that these proteins function in bacterial extracellular or periplasmic thiol-redox pathways.

MATERIALS AND METHODS

Sequence Similarity Searches and Multiple Sequence Alignments

We used the PSI-BLAST program[1] to search for homologs of a conserved group of hypothetical proteins belonging to an orthologous cluster (COG3019) described as "predicted metal-binding proteins" in the COG database (http://www.ncbi.nlm.nih.gov/COG).[17,18] Searches on the nr database (June 19, 2002: 1,012,231 sequences) with defined parameters (BLOSUM62 matrix, E-value threshold 0.02 or 0.05) were iterated to convergence starting with a single query sequence (gi|7517380). We grouped found homologs, using linkage clustering (score of 1 bit per site threshold, about 50% identity) as implemented in the SEALS package,[19] and used representative sequences from each group as new queries for subsequent rounds of PSI-BLAST. The iterations were repeated until no new sequences were detected. To retrieve additional COG3019 sequences, we also searched the unfinished microbial genomes database (http://www.ncbi.nlm.nih.gov/sutils/genom_table_cgi).

We constructed a multiple-sequence alignment for detected COG3019 homologs, using the program T-COFFEE.[20] Secondary structure predictions (JPRED server[21]) and hydrophobicity patterns guided manual adjustments to the alignment. We predicted signal sequences for each member sequence, using the SignalP server (http://www.cbs.dtu.dk/services/SignalP/); predicted cleavage sites based on either gram+ or gram− training sets were used to define the N-terminal boundaries of the respective sequences.[22,23] To detect remote homologs, a truncated COG3019 multiple sequence alignment lacking the signal sequence served as input to generate a position-specific scoring matrix or profile (-B option in blastpgp) for a new round of BLAST searches. We used each member sequence from the alignment as a query sequence to search the nr database using BLAST with the alignment-generated profile. Hits to these individual query sequences are

Fig. 1. Thioredoxin-like and predicted fold structure models and superposition. Ribbon diagrams representing (**A**) COG3019 rosetta model (DECOY_676), (**B**) bacterial disulfile oxidoreductase NrdH (1h75[49]), and (**C**) *E. coli* arsenate reductase ArsC (1j9b[37]) were produced with the program BOBSCRIPT.[50] Corresponding secondary structural elements are colored identically in rainbow from the N-terminus to the C-terminus of the thioredoxin-like fold. Elements corresponding to inserted domains are white. Residues conserved between all three groups, which are involved in disulfide exchange, are depicted as a large red ball-and-stick. Residues conserved among individual groups are depicted as a ball-and-stick. The orange sphere in ArsC represents a sulfate ion and depicts the active site. (**D**) Stereo diagrams of the backbone traces of (D) bacterial disulfide oxidoreductase NrdH (1h75, red) superimposed with the β-subdomain–truncated rosetta decoy structure (DECOY_1158, blue) and (**E**) the β-subdomain of *E. coli* ArsC (1j9b, red) superimposed with the β-subdomain of the COG3019 rosetta model (DECOY_676, blue) were generated with the program BOBSCRIPT.[50]

Fig. 2. Multiple-sequence alignment of thioredoxin-like domains. Each sequence is labeled according to NCBI gene identification (gi) number, PDB identifier, or Microbial Genome Database reference number. Sequences are grouped corresponding to SCOP families or COG (group I predicted metal binding protein or COG3019, group II thioltransferase or COG0695, and group III ArsC or COG1393). Sequence labels are colored black (bacterial), red (archaeal), green (viral), or blue (eukaryotic) according to taxonomy. The sequence identifier corresponding to the sequence used for generating the ab initio rosetta structure model is underlined. The first and last residue numbers are indicated before and after each sequence, with the total sequence length following in parentheses. Unconserved residues found between structural elements are omitted, with the number of missing residues in parentheses. Residues are highlighted according to hydrophobicity and size (large hydrophobic in yellow and small in gray), and conservation among groups (black). The secondary structures illustrated above the alignment correspond to β-strands (arrows) and α-helices (cylinders) found in the structures (1j9b and 1h75), and are colored in rainbow from N- to C-terminus of the Thioredoxin-like fold (see Fig. 1). The secondary structural elements (E for strand and H for helix) predicted by a program (Pred) used as a component of JPRED and the reliability of the prediction (rel) are shown below the alignment. Residues from sequences corresponding the BLAST hit between the thioltransferase group and COG3019 are highlighted in pink, with identities bolded.

reported with the BLAST statistics (E-value) produced by this procedure.

Sequences were grouped and classified according to the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop/i) and the COG database (http://www.ncbi.nlm.nih.gov/COG). Using various methods, we generated individual multiple-sequence alignments of families representing potential homologs to COG3019 (thioltransferase and ArsC) detected by BLAST methods. We generated an alignment of sequences corresponding to representative thioltransferase (group II) structures (1h25, 1dby, 1erv, 1kte, 1ilo, 1fo5, 1aba, 1qfn, and 1fov) using FSSP (http://www.ebi.ac.uk/dali/fssp/fssp.html)[24] and adjusted it after manual inspection. The thioltransferase sequences (gi|13473735 and gi|15805098) detected by similarity searches that do not have available structures were aligned based on BLAST alignments to the closest representative structure. We aligned representative ArsC (group I) sequences, including one structure (1j9b), using the program T-COFFEE,[20] and manually adjusted them based on hydrophobicity and secondary structure predictions (JPRED server[21]).

Tertiary Structure Prediction and Multiple Sequence Alignments

To accompany the BLAST statistics and further support the structure prediction of the target protein group, we applied fold-recognition (threading) methods to several member sequences. We submitted full-length sequences from gi|10047067, gi|7517380, and gi|18313562 to the hybrid fold-recognition method of Fischer[5] found on the BIOINBGU server (http://www.cs.bgu.ac.il/~bioinbgu/), which incorporates evolutionary information into a traditional threading procedure. We submitted the truncated multiple-sequence alignment lacking the signal sequence to the 3D-PSSM (three-dimensional, position-specific scoring matrix) server (http://www.sbg.bio.ic.ac.uk/~3dpssm/)6 and to the FUGUE server (http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html/).[7] Each of these methods combines multiple-sequence alignment profiles with 3D structure information to improve fold recognition.

We applied the rosetta ab initio protein structure prediction method[9] to the target sequence gi|11292822, with its N-terminus starting at the predicted signal sequence cleavage site (N-^{516}QAKEIQIY…GDKKL^{145}-C), and to a truncated version of the same sequence lacking its small C-terminal β subdomain (N-^{16}QAKEIQIY…NKPKD^{98}-C). The rosetta procedure first uses sequence profile comparisons to extract fragment libraries (for all possible 3- and 9-residue segments of the target chain) from the protein structure database. Using these libraries, rosetta applies Monte Carlo fragment substitution and optimization to construct a large number of independent 3D conformations.[12,13] For each target sequence, we generated 2000 independent fold decoys,[12,25] which were clustered based on RMSD.[9,14] The coordinates for the center decoys of the top clusters were submitted to the Dali (http://www.ebi.ac.uk/dali/) server[26,27] for structure comparison against the Protein Data Bank (PDB) database (Mon July 8, 2002: 3241 protein chains) and to generate structure-based alignments.

We merged the multiple alignments of each group into a global alignment using secondary structure predictions and hydrophobicity patterns, as previously described. Additionally, paired-BLAST hit alignments, fold-recognition structure–sequence alignments, and model-based structure–structure alignments guided the merging of the three multiple-sequence alignments.

RESULTS

Sequence Similarity Searches

The detection of remote homologs is often useful in predicting the structures and/or functions of new or uncharacterized protein families. In attempts to find remote homologs to an unclassified cluster of orthologous proteins (COG3019) described as predicted metal-binding proteins in the COG database,[17,18] we performed transitive PSI-BLAST searches starting from a member query sequence (gi|7517380). Using a standard E-value cutoff (0.02), we identified all members of COG3019 contained in the nr database (10 bacterial sequences and 1 archaeal sequence) and the unfinished microbial database (18 bacterial sequences). Upon construction of a multiple-sequence alignment of all found sequences, we noticed the presence of an N-terminal span of hydrophobic residues following one or two positively charged residues in each sequence (Fig. 2, group I). Although these N-terminal extensions do not display a large degree of conservation, they do exhibit the three distinct components found in signal peptides that mark proteins for secretion: a positively charged N-terminus; a long hydrophobic segment, including glycine or proline residues predicted to form an α-helix; and a C-terminal cleavage site marked by small and helix-breaking residues.[28] Submission of each sequence to the SignalP signal sequence server predicted the presence of export signal peptides and estimated the corresponding cleavage sites based on a combination of several artificial neural networks trained on signal peptides from either gram+ or gram− bacteria.[23]

Using a higher E-value cutoff (0.05), our transitive PSI-BLAST searches of the nr database linked this group of uncharacterized sequences with all families belonging to the thioredoxin-like fold superfamily (initial hit to gi|13473735 with query gi|7517380 in iteration 2, E-value 0.049). To further establish the family with the closest link to COG3019, we used the signal peptide-truncated COG3019 multiple-sequence alignment as a position-specific scoring matrix or profile for a new round of BLAST searches. With this procedure, all sequences present in the alignment find the thiotransferase-like sequence (gi|15805098), with the best hit (E-value 0.02) representing *Ralstonia metallidurans* CopG (gi|10047067). The BLAST alignment of this hit, which displays 27% sequence identity over a significant portion of the sequence (44 residues), is illustrated in Figure 2 (pink highlights). Additionally, all but two sequences find the hypothetical glutaredoxin-like sequence (gi|141400), albeit with higher E-values (lowest E-value 0.14). Each of these sequence hits

TABLE I. Similarity Search and Fold Recognition Results

Searching method	Query	Best hit	Score	Confidence[a]	SCOP family
PSI-BLAST (iterative)	gi|9625486	gi|13473735	E-value 0.049[b]	<0.02	Thioltransferase
PSI-BLAST (profile)	gi|10047067	gi|15805098	E-value 0.02	<0.02	Thioltransferase
Fischer	gi|7517380	1h75	Consensus 32.7	>12.0	Thioltransferase
Fischer	gi|10047067	1h75	Consensus 23.4	>12.0	Thioltransferase
Fischer	gi|18313562	1dby	Consensus 13.3	>12.0	Thioltransferase
3D-PSSM	Alignment	1dby	E-value 0.394[b]	<0.05	Thioltransferase
FUGUE	Alignment	1h75	Z score 7.09	>6.00	Thioltransferase

[a]Confidence thresholds for threading methods are gleaned from publications of the various methods: Fischer,[5] 3D-PSSM,[6] and FUGUE.[7]
[b]Scores do not fall within reported confidence thresholds.

belongs to either the glutaredoxin (COG0695) or the thioredoxin and thiol disulfide isomerase (COG0526) orthologous clusters, suggesting the SCOP-defined thioltranferase family (which includes both COGs) to contain the closest thioredoxin-like sequences to the uncharacterized protein cluster (results are summarized in Table I).

Fold Recognition

Given the modest BLAST statistics supporting the inclusion of COG3019 in the thioredoxin-like fold superfamily, we sought to substantiate our prediction using several fold-recognition (threading) methods. For different COG0319 sequences, the consensus fold-recognition method of Fischer (BIOINBGU server) that combines sequence, structural, and evolutionary information[5] produces hits to thioredoxin-like folds. Two sequences (gi|7517380 and gi|10047067) produce top hits to the bacterial oxidoreductase NrdH (1h75), with impressive consensus scores (32.7 and 23.4, respectively). The third sequence (gi|18313562) produces a top hit to chloroplast thioredoxin M (1dby, score 13.3). With the use of the COG0319 multiple-sequence alignment as input, results from 3D-PSSM provide all thioredoxin-like folds as hits, with the top three (1dby, E-value 0.394; 1fb6, E-value 0.444; 1h76, E-value 0.906) described as "worthy of attention."[6] Finally, profile hits to the COG0319 alignment with use of the FUGUE fold-recognition method belong to the thioredoxin-like fold superfamily, with the top hit (1h75, Z score 7.09) described as "certain" (Z score $\geq$6.0, 99% confidence) and the second hit (1a81, Z Score 4.53) described as "likely" (Z score $\geq$4.0, 95% confidence).[7] The predictions and scores of the top hits to each method are summarized in Table I.

Although the fold recognition methods all detect thioredoxin-like folds as hits, they produce different structure-based sequence alignments surrounding β-strand c and β-strand d. Such differences result in shifts of these structural elements and alternate placement of potential active site residues. For example, the 3D-PSSM alignment of the query sequence (gi|10047067) places the conserved Cys-His with Ile-Pro of the top hit (1dby) and splits the conserved Glu-Gly-His, with a three-residue gap between Glu and Gly. An alignment of the same query sequence with 1h75 places the Gly of the conserved Glu-Gly-His with that of 1h75 and splits the conserved Cys-His with a one-residue gap, placing the His with a Gln two residues upstream from the 1h75 Pro (Fig. 2). Thus, each of these alignments places the predicted β-strands c and d in different structural positions.

Rosetta Ab Initio Structural Model

The results of the sequence similarity searches, when combined with those of the fold-recognition programs, suggest an evolutionary link between the COG3019 family and the thioltransferase family. Accordingly, members of COG3019 should adopt a thioltransferase-like fold. To test this hypothesis independently and to differentiate between the various alignments produced by 3D-PSSM, we used the rosetta program to build a structural model of a COG3019 target sequence (gi|11292822). Starting with a truncated version of this sequence lacking the N-terminal signal peptide, we generated and clustered 2000 independent fold decoys (see Materials and Methods section). We considered the top hit to be the coordinates of the center (Decoy_1183) of the cluster containing the largest number of folds (50 decoys). The fold of this top hit partially resembles that of the thioredoxin-like superfamily, containing the four strands of the β-sheet and the first two helices (A and B) in roughly the correct orientations. However, the final helix C packs on the wrong side of the sheet against an unusual β-structure, in which β-strand c has more than two neighbor strands strands (b, d, and the sequence following helix C), as defined by Ruczinski et al.[29] Based on an analysis of existing β-sheet topologies, such an arrangement is not found in natural structures and should be filtered out when evaluating rosettta models.[29]

The structural model of the center (Decoy_676) of the next largest cluster (44 decoys) more closely resembles the thioredoxin-like fold [Fig. 1(A)]. In fact, use of the coordinates of this decoy to search the PDB database with Dali yields a top hit to protein disulfide oxidoreductase (1a8l, Z score 3.1, RMSD 3.7 over 90 residues). The rosetta model retains all of the structural elements of the thioredoxin-like fold, followed by an additional β-subdomain (d′ and d″). A similar β-subdomain is found following the ArsC arsenate reductase structure [Fig. 1(C)], which Dali aligns with an RMSD of 2.64 [Fig. 1(E)]. Structural comparisons of the remaining decoys of this cluster show that deviations tend to exist in three elements of the fold: helix B and the following loop; helix C; and the β-strand subdomain including d′ and d″. Interestingly, these areas tend to have

the greatest deviations in existing thioredoxin-like folds, and they are the areas in which insertions are prevalent.

One striking feature of the predicted decoy structure is the positioning of conserved residues in the thioredoxin-like fold active site. This active site, marked by the CYS-X-X-CYS motif [Cys residues illustrated in large red ball-and-stick, Fig. 1(A, B, and C)], is located along one edge of the β-sheet (predominantly C-terminal edge). Although the decoy contains the two conserved thiol-active cysteine residues, the conserved Pro residue found in the thioltransferases and the ArsC reductases is represented by a COG3019-specific His residue. In addition to this His residue, several other family-conserved residues putatively define the active site of the decoy structure [Cys, His, Ser, and Glu shown as ball-and-stick, Fig. 1(A)].

Because rosetta performs better on smaller proteins,[30] we decided to submit a truncated version of the target sequence lacking the extra β-subdomain [illustrated in white, Fig. 1(A)] to the rosetta algorithm. For this shortened target, the center decoys of the top two clusters (46 and 25 decoys, respectively) embody the thioredoxin-like fold. A Dali search of the structure database with the coordinates of the top decoy (decoy_1158) finds several thioredoxin-like superfamily members (Z scores 4.9): thioredoxin (1thx, RMS 3.2 over 66 residues), nitrogen regulation fragment Ure2P (1hqo, RMS 2.9 over 69 residues), glutaredoxin-like protein NrdH (1h75, RMS 3.5 over 68 residues), and bacteriophage T4 glutaredoxin (1aba, RMS 2.7 over 64 residues). The remaining significant hits all belong to the thioredoxin-like fold superfamily (1gwc, 1gnw, 1g7o, 1kte, 1erv, 1k0n, 1qfn). A stereo diagram of the superposition of the decoy_1158 trace [blue, Fig. 1(D)] with the 1h75 trace [red, Fig. 1(D)] shows the extent of the similarity of the predicted model with the thioredoxin-like fold.

Multiple-Sequence Alignment

The model predicted by the rosetta procedure allows us to structurally align the COG3019 multiple-sequence alignment with other thioredoxin-like family multiple-sequence alignments, and helps differentiate between various alignments produced by traditional fold-recognition methods. To complete this task, we chose representative sequences of two thioredoxin-like families: the thioltransferases (Fig. 2, Group II), which include sequences detected in sequence similarity searches with COG3019, and the ArsC reductases (Fig. 2, Group III), which contain additional family-specific residue conservations that mark the thioredoxin-like fold active site. Due to the presence of a wide variety of thioltransferase structures in the PDB, we derived a multiple-sequence alignment of this family based on Dali structural superpositions. The sequences (gi|15805098 and gi|13473735) detected in similarity searches that do not have available structural data were included in the multiple-sequence alignment based on paired BLAST hit alignments to the closest representative thioltransferase structures (Fig. 2, Group II). We aligned ArsC sequences using the program T-COFFEE and manually adjusted the alignment based on hydrophobicity and secondary structure predictions (Fig. 2, Group III).

Comparisons of the three individual, multiple-sequence alignments revealed several conserved characteristics. The three alignments display similar patterns of hydrophobicity, produce comparable secondary structure predictions, share conserved thiol-redox active cysteines of the motif C-X-X-C, and retain a conservation of residue conservations within the alignments (Fig. 2, black highlights). We therefore used these patterns and conservations, along with paired-BLAST hit alignments, fold-recognition structure–sequence alignments, and model-based structure–structure alignments to guide merging the three multiple-sequence alignments into the global multiple-sequence alignment illustrated in Figure 2.

DISCUSSION

Validity of Fold Prediction and Structural Model

We infer homology between the uncharacterized COG3019 protein family and the thioltransferase family based on several lines of evidence. First, the statistics produced by various PSI-BLAST searching strategies support the proposed evolutionary link. Although transitive PSI-BLAST searches starting from a COG3019 query sequence (gi|7517380) detect a thioredoxin sequence intermediate sequence (gi|13473735) with a marginal E-value (0.049), these statistics improve (E-value 0.02) with the use of an alignment of all detected COG3019 sequences as a profile for PSI-BLAST (Table I). Second, the structure-based fold-recognition methods we chose to evaluate our queries all take into account evolutionary relatedness. Each of these methods compares query sequence profiles derived from PSI-BLAST or generated from an input alignment to structure profiles derived by various approaches. The hybrid fold-recognition of Fischer considers both the single-structure sequence and the PSI-BLAST structure-alignment profile in its consensus score,[5] whereas FUGUE generates profiles from alignments of homologous structures.[7] The 3D-PSSM method obtains a profile based on both PSI-BLAST–derived alignments and structural homolog–derived alignments.[6] With one exception (3D-PSSM score), scores for all of the COG3019 structure predictions fall at or within the confidence limits of the respective methods (Table I).

Althought these homology-based methods detect thioltransferases as significant hits to COG3019 sequences, the assembly of a global multiple-sequence alignment, this superfamily remains challenging because of the significant divergence of the sequences and structures within the fold group. Many thioredoxin-like structures are placed within their own groups in FSSP, having low sequence identities (below 10%), and containing large insertions within the common fold. For structures belonging to the same SCOP family (thioltransferases), FSSP assigns Z scores as low as 4.7. These scores become even lower when structures belong to different thioredoxin-like fold families (i.e., thioltransferases vs protein disulfide isomerases, with Z scores as low as 2.7). Accordingly, PSI-BLAST generally detects only structural elements surrounding

the conserved CYS-X-X-CYS motif for this superfamily, limiting alignments to the first half of the sequence. The 3D-PSSM fold-recognition method also provides conflicting results in the alignment of COG3019 sequences with the major structural elements (i.e., β-strands c and d) of the thioredoxin-like fold. Although this method considers multiple structures in its predictions, it groups only similar thioredoxin-like structures (less than 6A RMSD) in its fold library, potentially losing structural information from this diverse superfamily. Additionally, regular patterns of hydrophobicity are not very well maintained in the thioredoxin-like sequences (as illustrated by a lack of yellow highlights, or differing yellow highlights between families in Fig. 2), superseding the use of solvation potentials in the 3D-PSSM predictions.

Such sequence and structural variation between the thioredoxin-like folds may indicate that evolutionary-based sequence alignments vary from structure-based sequence alignments for this protein superfamily. In such cases, structural information becomes increasingly important in the assembly of multiple-sequence alignments. Consequently, we took advantage of the relatively large number of diverse structures belonging to the thioltransferase and ArsC families, using structure-based alignments of the different sequences to guide our multiple-sequence alignment. In addition to providing an independent conformation of homology-based COG3019 fold predictions, the rosetta model helps to guide the merging of thioredoxin-like multiple-sequence alignments into a global alignment. As illustrated in Figure 1, the rosetta model closely resembles thioltransferase structures, displaying a convincing turn conformation (between β-strands c and d) and maintaining conserved residues in positions that make up the active sites of other thioredoxin-like folds. We therefore used comparisons of this model with other thioltransferase structures to distinguish between the various alignments produced by the 3D-PSSM fold recognition method. The resulting structural alignment maintains a conservation of conservations (conserved thioltransferase Pro corresponding to conserved COG3019 His and conserved ArsC Args corresponding to conserved COG3019 Cys and His; black highlights in Fig. 2) and places gaps in similar regions (between β-strands c and d). Thus, we combine a variety of homology- and structure-based fold prediction methods to identify COG3019 as a thioltransferase and to produce a structural model of the thioltransferase-like domain of this previously uncharacterized protein family.

The rosetta structure model contains a β-subdomain following the core thioredoxin-like fold of the COG3019 sequences. Despite the structural similarity of this predicted subdomain with that found in the ArsC structure [Fig. 1(E)], many ArsC sequences end prior to these two β-strands. Such a loss suggests a lack of functional or structural significance of this subdomain within the ArsC family. For the COG3019 sequences, this subdomain is also not very well conserved. Secondary structure predictions for some COG3019 sequences (gi|18313562) suggest the presence of both β-strands, whereas predictions for other sequences (gi|10047067 and DOE_48935) suggest the presence of only one strand. When combined with the limited length of the sequence corresponding to this structural segment, the lack of conservation makes an assessment of homology difficult. We therefore omit this section of the ArsC and COG3019 sequences from the global multiple-sequence alignment, limiting it to the structural elements of the core thioredoxin-like fold.

Functional Implications of Fold Prediction

Based on sequence similarity, fold recognition, and ab initio structure prediction, sequences belonging to the previously uncharacterized COG3019 family are predicted to adopt a thioredoxin-like fold. Their close evolutionary link to members of thioltransferase family also suggests a functional prediction for the COG3019 proteins. The active-site sequence motif CYS-X-X-CYS is present in all COG3019 sequences as CYS-GLY-CYS-CYS. For the thioltransferases thioredoxin and glutaredoxin, which generally function as cytoplasmic thiol-disulfide reductants, the residues belonging to this sequence motif play an important functional role in thiol-redox chemistry. An increase in the reducing ability of these enzymes correlates with increase in the pKa of the first Cys residue of this motif (or a relative stabilization of the oxidized form of the enzyme), which is profoundly influenced by the the nature of the remaining residues.[31–33] Accordingly, the presence of the Gly-Cys dipeptide between the potential disulfide pair of the COG3019 sequences may influence the redox potential of these enzymes and should determine their general function as oxidants or reductants.

The thiol-disulfide oxidoreductase families contain another conserved residue (pro) in close structural proximity to the active-site motif (C-X-X-C). This residue, which exists in the *cis*-conformation, plays a major role in the folding and stability of the enzyme,[16,34] and its mutation in one thiol-disulfide oxidase (DsbA) also influences the pKa of the redox active Cys.[16] Although this *cis*-Pro is conserved in most thioredoxin-like folds, the corresponding residues (Glu and Gln) in two thioredoxin-like domains of calsequestrin and the residue (Lys) in the thioredoxin-like domain of phenol hydroxylase do not form *cis*-peptide bonds, suggesting that this configuration is not necessary for all thioredoxin-like folds. In the COG3019 sequences, the Pro is replaced by a conserved His residue. Like the thioltransferases, this His could potentially form a *cis*-peptide bond. Indeed, nonproline *cis*-peptide bonds occur in protein structures,[35,36] with one example (trimethylamine dehydrogenase, 2tmd) containing a His residue in the *cis* conformation.[35]

The COG3019 sequences contain additional conserved residues (His, Cys, and Glu) positioned in the model structure to form a potential active site. In the structure of ArsC, the corresponding active-site residues form hydrogen bonds with a thiolate-bound arsenic substrate (arg-107) and a bisulfite anion that mimics the product (arg-94) and may participate in the reaction mechanism.[37] Accordingly, the conserved COG3019 residues may participate in substrate binding and/or reaction chemistry. Alterna-

tively, these residues could bind a metal ion, as suggested by the "predicted metal-binding protein" description of the COG3019 members in the COG database.[17,18] Perhaps metal binding could provide a switch that regulates redox potential, as is seen in the thiol-disulfide redox switch formed between three essental Cys residues of the zinc-binding anti-sigma factor (RsrA)[38,39] and the oxidation-induced chaperone activity of heat shock protein HSP33, which is also mediated through redox-sensitive Cys coordination of zinc.[40] The C-G-C-C motif of the COG3019 sequences contains adjacent Cys residues. Disulfide bonds can exist between adjacent Cys residues in proteins, and their reduction can control protein function and influence conformational stability.[41–43] Such a motif has even been engineered as a redox switch to control the activity of ribonuclease A.[44] Thus, the conserved residues of the COG3019 family could all potentially coordinate metal ions. The functional clustering of several COG3019 members found on the ERGO database (http://ergo.integratedgenomics.com/ERGO/) with multicopper oxidases, copper transporting ATPases, and cobalt-zinc-cadmium resistance proteins (czcA and czcB) also supports their association with metals.[45]

The presence of signal peptide in the COG3019 sequences is suggestive of their export to the bacterial periplasm or extracellular space, where they may function in thiol-redox pathways. An active thiol isomerization system exists in the bacterial periplasm, including several enzymes with thioredoxin-like folds (DsbA, DsbC, and others). These proteins function as either reductases (DsbD and DsbE) or oxidases (DsbA and DsbB) to maintain disulfide bonds in newly formed and translocated polypeptides, and are thus important for extracellular protein folding and function.[46,47] Additionally, several of these disulfide oxidoreductases have been linked to C-type cytochrome biogenesis, suggesting a role in electron transport of the respiratory chain.[48] Although the precise COG3019 function in any of these known pathways or in a novel thiol-redox pathway has yet to be determined, the homology-based predictions presented in this article provide the initial framework for such experimental conformation.

REFERENCES

1. 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.
2. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1999;27:260–262.
3. Gerstein M. Measurement of the effectiveness of transitive sequence comparison, through a third "intermediate" sequence. Bioinformatics 1998;14:707–714.
4. Kinch LN, Grishin NV. Expanding the nitrogen regulatory protein superfamily: Homology detection at below random sequence identity. Proteins 2002;48:75–84.
5. Fischer D. Hybrid fold recognition, combining sequence derived properties with evolutionary imformation. In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE, editors. Pacific Symposium on Biocomputing. Oahu, Hawaii: World Scientific; 2000. p 119–130.
6. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
7. Shi J, Blundell TL, Mizuguchi K. FUGUE: Sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310:243–257.
8. Lesk AM, Lo Conte L, Hubbard TJ. Assessment of novel fold targets in CASP4: Predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins 2001;45(Suppl 5):98–118.
9. Bonneau R, Tsai J, Ruczinski I, Chivan D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: Progress in ab initio protein structure prediction. Proteins 2001;45(Suppl 5):119–126.
10. Han KF, Bystroff C, Baker D. Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. Protein Sci 1997;6:1587–1590.
11. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci U S A 1996;93:5814–5818.
12. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
13. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
14. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci U S A 1998;95:11158–11162.
15. 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.
16. Charbonnier JB, Belin P, Moutiez M, Stura EA, Quemeneur E. On the role of the cis-proline residue in the active site of DsbA. Protein Sci 1999;8:96–105.
17. Tatusov RL, Natale DA, Garkautsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryotin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: New developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29:22–28.
18. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000;28:33-36.
19. Walker DR, Koonin EV. SEALS: A system for easy analysis of lots of sequences. Proc Int Conf Intell Syst Mol Biol 1997;5:333–339.
20. Notredame C, Higgins DG, Heringa J. T-COFFEE: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
21. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: A consensus secondary structure prediction server. Bioinformatics 1998;14:892–893.
22. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997;8:581–599.
23. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
24. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–603.
25. Bonneau R, Strauss CE, Baker D. Improving the performance of rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 2001;43:1–11.
26. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16:566–567.
27. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
28. Izard JW, Kendall DA. Signal peptides: Exquisitely designed transport promoters. Mol Microbiol 1994;13:765–773.
29. Ruczinski I, Kooperberg C, Bonneau R, Baker D. Distributions of beta sheets in proteins with application to structure prediction. Proteins 2002;48:85–97.
30. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.
31. Grauschopf U, Winther JR, Korber P, Zander T, Dallinger P, Bardwell JC. Why is DsbA such an oxidizing disulfide catalyst? Cell 1995;83:947–955.
32. Huber-Wunderlich M, Glockshuber R. A single dipeptide sequence

modulates the redox properties of a whole enzyme family. Fold Des 1998;3:161–171.
33. Guddat LW, Bardwell JC, Glockshuber R, Huber-Wunderlich M, Zander T, Martin JL. Structural analysis of three His32 mutants of DsbA: Support for an electrostatic role of His32 in DsbA stability. Protein Sci 1997;6:1893–1900.
34. Kelley RF, Richards FM. Replacement of proline-76 with alanine eliminates the slowest kinetic phase in thioredoxin folding. Biochemistry 1987;26:6765–6774.
35. Jabs A, Weiss MS, Hilgenfeld R. Non-proline cis peptide bonds in proteins. J Mol Biol 1999;286:291–304.
36. Pal D, Chakrabarti P. Cis peptide bonds in proteins: Residues involved, their conformations, interactions and locations. J Mol Biol 1999;294:271–288.
37. Martin P, Demel S, Shi J, Gladysheva T, Gatti DL, Rosen BP, Edwards BF. Insights into the structure, solvation, and mechanism of ArsC arsenate reductase, a novel arsenic detoxification enzyme. Structure 2001;9:1071–1081.
38. Paget MS, Bae JB, Hahn MY, Li W, Kleanthous C, Roe JH, Buttner MJ. Mutational analysis of RsrA, a zinc-binding anti-sigma factor with a thiol-disulphide redox switch. Mol Microbiol 2001;39:1036–1047.
39. Kang JG, Paget MS, Seok YJ, Hahn MY, Bae JB, Hahn JS, Kleanthous C, Buttner MJ, Roe JH. RsrA, an anti-sigma factor regulated by redox change. EMBO J 1999;18:4292–4298.
40. Jakob U, Eser M, Bardwell JC. Redox switch of hsp33 has a novel zinc-binding motif. J Biol Chem 2000;275:38302–38310.
41. Kim BM, Schultz LW, Raines RT. Variants of ribonuclease inhibitor that resist oxidation. Protein Sci 1999;8:430–434.
42. Blake CC, Ghosh M, Harlos K, Avezoux A, Anthony C. The active site of methanol dehydrogenase contains a disulphide bridge between adjacent cysteine residues. Nat Struct Biol 1994;1:102–105.
43. Miller SM, Moore MJ, Massey V, Williams CH Jr, Distefino MD, Ballou DP, Walsh CT. Evidence for the participation of Cys558 and Cys559 at the active site of mercuric reductase. Biochemistry 1989;28:1194–1205.
44. Park C, Raines RT. Adjacent cysteine residues as a redox switch. Protein Eng 2001;14:939–942.
45. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 1999;96:2896–2901.
46. Raina S, Missiakas D. Making and breaking disulfide bonds. Annu Rev Microbiol 1997;51:179–202.
47. Ritz D, Beckwith J. Roles of thiol-redox pathways in bacteria. Annu Rev Microbiol 2001;55:21–48.
48. Thony-Meyer L. Biogenesis of respiratory cytochromes in bacteria. Microbiol Mol Biol Rev 1997;61:337–376.
49. Stehr M, Schneider G, Aslund F, Holmgren A, Lindqvist Y. Structural basis for the thioredoxin-like activity profile of the glutaredoxin-like NrdH-redoxin from *Escherichia coli*. J Biol Chem 2001;276:35836–35841.
50. Esnouf RM. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. J Mol Graph Model 1997;15:132–134, 112–113.