Record, M. T., Anderson, C. F., & Lohman, T. L. (1978) *Q. Rev. Biophys. 11*, 103.
Santoro, M. M., & Bolen, D. W. (1988) *Biochemistry 27*, 8063.
Sato, K., & Egami, F. (1957) *J. Biochem. (Tokyo) 44*, 753.
Schellman, J. A. (1975) *Biopolymers 14*, 999.
Shirley, B. A., & Laurents, D. V. (1990) *J. Biochem. Biophys. Methods 20*, 181.
Shirley, B. A., Stanssens, P., Steyaert, J., & Pace, C. N. (1989) *J. Biol. Chem. 264*, 11621.
Shirley, B. A., Stanssens, P., Hahn, U., & Pace, C. N. (1992) *Biochemistry 31*, 725.
Sturtevant, J. M. (1987) *Annu. Rev. Phys. Chem. 38*, 463.
Takahashi, K., Uchida, T., & Egami, F. (1970) *Adv. Biophys. 1*, 53.
Tanford, C. (1961) *Physical Chemistry of Macromolecules*, p 554, John Wiley & Sons, New York.
Tanford, C. (1968) *Adv. Protein Chem. 23*, 121.
Thomson, J. A., Shirley, B. A., Grimsley, G. R., & Pace, C. N. (1989) *J. Biol. Chem. 264*, 11614.
Walz, F. G., & Kitareewan, S. (1990) *J. Biol. Chem. 265*, 7127.
Xie, D., Bhakuni, V., & Freire, E. (1991) *Biochemistry 30*, 10673.

Identification of the Three-Dimensional Thioredoxin Motif: Related Structure in the ORF3 Protein of the *Staphylococcus aureus mer* Operon†

Lynda B. M. Ellis,*,‡ Peter Saurugger,§ and Clare Woodward‖

Department of Laboratory Medicine and Pathology, Box 511 UMHC, University of Minnesota, Minneapolis, Minnesota 55455, and Molecular Biology Computing Center and Department of Biochemistry, University of Minnesota, St. Paul, Minnesota 55108

Received July 16, 1991; Revised Manuscript Received March 6, 1992

ABSTRACT: We have developed a computerized search pattern for recognition of the three-dimensional redox site of thioredoxins based on primary and predicted secondary structure. This pattern, developed in the ARIADNE protein expert system, is used to search for thioredoxin-like tertiary structural motif among proteins for which the only structural information is the primary sequence. The pattern was trained on 102 protein sequences (25 functionals and 77 controls); it matches all 25 members of the functional set under cutoff conditions that include only 2 members of the control set, for a sensitivity of 1.0 and a specificity of 0.97. The pattern matches only one of the two thioredoxin-like domains in protein disulfide isomerases (PDIs) and their analogues, suggesting that the C-terminal domain is more structurally similar to thioredoxin than the N-terminal domain. The *Escherichia coli* DsbA protein, a possible PDI analogue, appears to be more structurally similar to the N-terminal thioredoxin-like domain of PDIs. Thioredoxin-like redox functionality has been proposed for lutropin and follitropin, in part on the basis of their having -Cys-X-Pro-Cys- sequences. None match our pattern; all lack a predicted α-helix pattern element immediately after the active site. Hypothetical proteins in the National Biomedical Research Foundation Protein Identification Resource database were searched for matches to the pattern. The most interesting match was a hypothetical protein (161 residues) from the third open reading frame in the *Staphylococcus aureus mer* operon, which is involved in mercury detoxification. The match to our pattern and the hydrophobicity distribution in aligned elements of secondary structure not in our pattern strongly suggest that it has thioredoxin-like structure.

Prediction of tertiary structure and biological function from primary sequence alone is a central challenge in protein chemistry. Experimental determination of tertiary structure by X-ray crystallography or NMR requires months, while automated protein and gene sequencing techniques are increasingly efficient. This results in an ever-widening gap between the number of known protein primary sequences and the number of known three-dimensional structures. For example, while there are 39 533 loci (potential primary sequences) in GenBank[1] 65.0, as of January 1991 there were only 622 structures in the Protein Data Bank, and many of the latter are redundant. Further, the number of hypothetical proteins predicted from nucleotide sequences is increasing faster than that of proteins with known function.

To address the primary sequence–tertiary structure question, computer-based tools are being developed in many labs to correlate and recognize the tertiary structural information which resides in protein sequences. One common approach to prediction of the tertiary structure of a sequence is to search for primary sequence similarity to one or more proteins with known structure. When such similarity is found, the unknown protein can be aligned with the protein(s) of known structure.

† This work was supported in part by grants from the University of Minnesota Industry–University Cooperative Research Center for Biocatalytic Processing and the University of Minnesota Graduate School and by grants of computer time from the University of Minnesota Molecular Biology Computer Center.

‡ Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis.
§ Molecular Biology Computing Center, University of Minnesota, St. Paul.
‖ Department of Biochemistry, University of Minnesota, St. Paul.

[1] Abbreviations: GenBank, NIH nucleic acid sequence data bank; PIR, protein identification resource data bank; Swiss-Prot, University of Geneva protein sequence data bank; EMBL, European Molecular Biology Laboratory; PDI, protein disulfide isomerase; ORF, open reading frame.

0006-2960/92/0431-4882$03.00/0

This alignment is used to locate the secondary structural elements in the unknown protein, and to model its tertiary structure by homology to the known protein [for example, see Sali et al. (1990) and Sander and Schneider (1991)]. Benner and Gerloff (1990) have developed a powerful method to predict secondary structure that identifies insertion/deletion sites from numerous related primary sequences, and does not require a known tertiary structure.

There are methods for recognition of three-dimensional similarities in the case of very weak primary sequence similarity. Pattern matching using patterns more complex than primary sequence is one such approach, and the one utilized here. Other examples of this type of pattern matching include analysis of amino acid sequence patterns to predict secondary and supersecondary structural elements (Lim, 1974; Cohen et al., 1986; Edwards, 1987), to detect tertiary structural similarities (Taylor, 1986; Barton & Sternberg, 1990), and to identify determinants of a protein fold (Bashford et al., 1987; Ouzounis & Sander, 1991; Bowie et al., 1991; Finkelstein & Reva, 1991). Recent reviews include Taylor (1988), Fasman (1989), and Thornton et al. (1991).

The approach in ARIADNE, used here, is to identify a pattern of sequence elements and associated properties that is diagnostic of a set of proteins thought to contain a common structural motif (Webster et al., 1987; Lathrop et al., 1987). A motif is a unit of structure, usually smaller than a cooperative folding domain, and often containing the binding site of a functionally important ligand. Patterns are developed on a set of related, but not identical, proteins and then are used to search protein sequences for similar patterns indicating probable structural similarity. One important feature of ARIADNE is that its patterns can include both primary and predicted secondary structure. Here we report the development of a sequence pattern for recognition of the tertiary structural motif of the active site of a functionally important and ubiquitous protein family: the disulfide/dithiol oxidoreductases.

Thioredoxin and Related Disulfide Redox Proteins. The most well-studied disulfide oxidoreductases are the thioredoxins. Thioredoxins constitute a category of protein which is especially useful for pattern development because sequences are available from diverse phyla (animal, plant, bacteria, bacteriophage), at least one representative X-ray structure is known, and they form a part of the larger family of disulfide redox proteins with which they share very little sequence identity. Here we develop the basic thioredoxin motif pattern. It can now be refined and periodically applied to new sequences as they become available.

Thioredoxin is a small, soluble protein found in all organisms. Its redox functionality is due to an active-site disulfide/dithiol formed by the cysteines in the conserved sequence -Cys-Gly-Pro-Cys-. This sequence is located at the N-terminus of the long helix in a β–α–β structural unit. The reference structure for thioredoxins is the crystal structure of the 108-residue protein from *Escherichia coli* (Holmgren et al., 1975; Katti et al., 1990) composed of a central β-sheet of 5 strands packed by 4 helices, with the redox-active disulfide in a reverse turn between β_2 and α_2. A ribbon drawing of the canonical tertiary structure of thioredoxin is shown in Figure 1.

As protein disulfide reductases, thioredoxins are involved in a wide assortment of regulatory functions [for reviews, see Holmgren (1985) and Gleason and Holmgren (1988)]. These include regulation of the light and dark cycle of photosynthesis (Wolosiuk & Buchanan, 1977; Schürmann et al., 1981), the glucocorticoid receptor (Grippo et al., 1985), CMP kinase

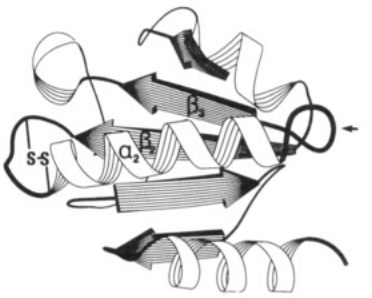

FIGURE 1: Oxidized *E. coli* thioredoxin. The ribbon drawing shows the five-stranded β-sheet and four helices. The active-site disulfide, β_2 and β_3 strands, and α_2-helix are labeled. The arrow indicates the α_2-β_3 loop. Adapted from Langsetmo et al. (1991).

(Maness & Orengo, 1975), the initiation of protein synthesis (Hunt et al., 1983), and the deiodination of thyroxine (Das et al., 1988).

In addition to its numerous redox functions, thioredoxin has been recruited for participation in the replication cycle of some bacteriophages. DNA polymerase in phage T7 forms a required complex with thioredoxin (Modrich & Richardson, 1975; Nordström et al., 1981; Huber et al., 1986). Assembly of filamentous phages M-13 and f1 also requires host thioredoxin (Russel & Model, 1985, 1986; Lim et al., 1985). Apparently, only the reduced form of thioredoxin is active in these systems, and the disulfide/dithiol redox activity is not required.

Glutaredoxins are a class of redox proteins related to thioredoxins. They have a redox-active disulfide bond formed by cysteines in the sequence -Cys-Pro-Tyr-Cys-. Glutaredoxins occur widely, and have been isolated from several microbial and mammalian species. Thioredoxins differ from glutaredoxins in the nature of their hydrogen acceptor. Thioredoxin donates hydrogens to several reductases, including thioredoxin reductase, while glutaredoxin donates hydrogens to glutathione. The functional and sequence properties of thioredoxins and glutaredoxins have been compared (Eklund et al., 1984, 1991; Gleason & Holmgren, 1988; Fuchs, 1989; Holmgren, 1989).

The reference structure for the glutaredoxins is the 85-residue T4 phage thioredoxin (Söderberg et al., 1978). T4 thioredoxin is more closely related to the glutaredoxins than to other thioredoxins (Eklund et al., 1984) and is better thought of as a glutaredoxin; its renaming to T4 glutaredoxin has been suggested (H. Eklund, personal communication). Thioredoxin and glutaredoxin structures differ most notably in that glutaredoxins lack the first N-terminal helix and strand found in thioredoxins. They also differ in size of various loops, and in a buried aspartic acid in the thioredoxins, which is an aliphatic side chain in the glutaredoxins. In *E. coli* thioredoxin, the buried aspartic acid is at position 26, and its pK_a of 7.5 is highly anomalous (Langsetmo et al., 1991). In both the thioredoxins and the glutaredoxins, there is a conserved proline (position 76 in the *E. coli* protein) which is in a cis peptide bond in the crystal structures. *E. coli* thioredoxin can also fold with the Pro-76 peptide bond in the trans form, but this configuration is less stable, especially in the oxidized (disulfide) protein (Langsetmo et al., 1989).

Another class of disulfide/dithiol redox enzymes related to thioredoxins are the protein disulfide isomerases (Edman et al., 1985). Protein disulfide isomerases (PDIs) are a group of large redox proteins (about 450 residues long) which contain 2 domains having high sequence similarity to thioredoxin, including the -Cys-X-X-Cys- active site (Freedman et al., 1989). A PDI may function as a disulfide isomerase, a proline hydroxylase (Pihlajaniemi et al., 1987), or a thyroxine de-

iodinase (Boado et al., 1988), and they are rightly called multifunctional proteins. In addition, there are several PDI analogues which also contain two domains with sequence similarities to thioredoxin. PDI analogues include proteins with functions as diverse as thyroid binding protein (Cheng et al., 1987), phosphoinositol–phospholipase C (Bennett et al., 1988), and glycosylation site binding protein (Geetha-Habib et al., 1988). Another PDI analogue is a nonvertebrate protein of unknown function from *Trypanosoma brucei* (Hsu et al., 1989). A recent interesting PDI analogue, or alternatively a new class of disulfide oxidoreductase proteins, has been reported (Bardwell et al., 1991). This DsbA protein is coded by the *dsbA* gene in *E. coli*, has 208 amino acid residues, contains the active-site sequence -Cys-Pro-His-Cys-, and is capable of reducing protein disulfide bonds in vitro and in vivo (Bardwell et al., 1991).

Thioredoxin Motif. Our thioredoxin motif pattern, or descriptor, is defined in terms of primary sequence and predicted secondary structure in the language of ARIADNE. ARIADNE is a Lisp-based expert system that supports user-directed pattern design, modification, and refinement (Webster et al., 1987). To be useful in this context, secondary structure predictions need not be highly accurate, as long as they overpredict actual secondary structure. Our descriptor was refined to give high sensitivity and specificity on a training set composed of a functional set to be matched and a control set to be excluded. The functional set includes the thioredoxins and related proteins, and the control set includes proteins containing -Cys-X-X-Cys-, but known to be unrelated to thioredoxin.

Our pattern has a very high degree of sensitivity and specificity for the thioredoxin redox motif. We have used this pattern to search the PIR database. The most interesting match is the hypothetical protein encoded by ORF3 in the *mer* operon in *Staphylococcus aureus*. We predict thioredoxin-like tertiary structure for this putative gene product. Also, we find that in PDIs the two thioredoxin-like domains are not equivalent. The DsbA protein coded by the *dsbA* gene in *E. coli* matches the PDI domain which is less structurally similar to thioredoxin.

Materials and Methods

The primary computational resource used in this study was the Molecular Biology Computer Center (MBCC) in the College of Biological Sciences, University of Minnesota. Preliminary work was carried out while one author (L.B.M.E.) was a visiting scientist at the Molecular Biology Computer Research Resource (MBCRR) of the Dana Farber Cancer Institute and Harvard University.

The PIR 25 and 26 and Swiss-Prot 14 protein databases and the EMBL 24.65 nucleic acid database were available as part of the Intelligenetics Suite 5.37 and 5.4 at the MBCC. These databases were searched using the Intelligenetics modules FINDSEQ (when loci names or keywords were known) or QUEST (for searches based on sequence patterns). PIR 26 was also accessed through the MBCRR or the University of Houston Gene-Server (Davison & Chapplear, 1990).

Several software tools in the MBCRR Package, available from the MBCRR, were used in this study. Truncation of sequences on either side of known primary sequence patterns was carried out using a simple regular expression tool, GGREP.FS. Clustering of sequences based on their primary sequence similarity was carried out with PIMA (Smith & Smith, 1990, 1991). Alignments of sequences was carried out with PIMA and MASE (multi-alignment sequence editor) (Faulkner & Jurka, 1988). Secondary structure prediction was carried out by PRSTRC (Ralph et al., 1987), based on a modification of the Chou–Fasman algorithm (Chou & Fasman, 1978). Patterns involving secondary and primary structure were matched against sequences using the ARIADNE expert system (Webster et al., 1987). Clusters were displayed in tree form using an MBCC tree-display program.

Table I: Functional Set in Reverse Similarity Cluster Order

PIR or Swiss-Prot locus name	title
THIM$ANANI	thioredoxin M, *Anacystic nidulans*
THIM$ANASP	thioredoxin M, Anabaena
THIM$SPIOL	thioredoxin M, spinach chloroplast
THI1$CORNE	thioredoxin C-1, *Corynebacterium nephridii*
THIO$ECOLI	thioredoxin, *E. coli*
THI2$CORNE	thioredoxin C-2, *Corynebacterium nephridii*
THIO$HUMAN	thioredoxin, human
THIO$RABIT	thioredoxin, rabbit
THIO$MOUSE	thioredoxin, mouse
THIO$CHICK	thioredoxin, chicken
THIF$SPIOL	thioredoxin F precursor, spinach chloroplast
PDI$MOUSE	PDI, mouse
S01634	thyroid hormone binding protein precursor, mouse
PDI$RAT	PDI, rat
PDI$BOVIN	PDI, bovine
PDI$HUMAN	PDI, human
A30007	glycosylation site binding protein, chicken
PDI$CHICK	PDI, chicken
A28807	phosphoinositol–phospholipase C, form I, rat
BS2$TRYBR	bloodstream protein 2 precursor, *Trypanosoma brucei*
GLRX$BOVIN	glutaredoxin, bovine
GLRX$PIG	glutaredoxin, pig
GLRX$RABIT	glutaredoxin, rabbit
GLRX$ECOLI	glutaredoxin, *E. coli*
THIO$BPT4	glutaredoxin, bacteriophage T4

Results

Functional Set. The PIR and Swiss-Prot protein and EMBL DNA databases were searched for sequences of known thioredoxins, PDIs and their analogues, and glutaredoxins. Twenty-five sequences (11 thioredoxins, 9 PDIs and analogues, and 5 glutaredoxins) were the functional set for this study. Loci names and sequence titles are given in Table I. While more sequences were found, some were duplicates or precursors of those included. Others contained ambiguous residue codes (X, Z, or B) and were excluded since these could not be used by the secondary structure prediction software PRSTRC. The sequence for spinach M thioredoxin found in Swiss-Pot was corrected (Eklund et al., 1991).

This functional set of 25 sequences was aligned by the PIMA sequence alignment software (Smith & Smith, 1990). PIMA also compares each pair of sequences, assigns each pair a similarity score, and on the basis of this score assigns sequences to similarity clusters and subclusters as illustrated in Figure 2. Those protein pairs with highest similarity scores (joined lower on the ordinate) are most similar. (In Figure 2, these are the PDIs.) To develop the initial screening pattern, we identified 10 sequences from Figure 2 which represent the sequence diversity of the entire functional set. The common primary sequence attributes of the functional set are

pattern 1

[VIL]..?[FYW]....?.?.?C..C

where the brackets enclose residue alternatives for a given position, a dot indicates any residue is acceptable for a position, and a question mark indicates one or zero instances of the preceding residue. Pattern 1 is read as follows: a hydrophobic aliphatic residue (V, I, or L) separated by one or two residues from an aromatic residue (F, Y, or W) separated by three to

Table II: Control Set: PIR Locus Name, Size of Cluster Each Represents, and Title

name	size	title
FESGAL	52	ferredoxin, *Spirulina platensis*
DEHOAL	43	alcohol dehydrogenase E chain, horse
A05106	32	protein kinase C II, rat
A24628	29	collagen α 2(IV) chain, mouse (fragment)
S06592	27	phosphoprotein phosphatase 2A-α, catalytic chain, rat
A32224	16	androgen receptor, human
TIQJM	17	ovomucoid (PSTI-type protease inhibitor) 1, Japanese quail
TVVPAS	16	large T antigen, polyomavirus BK (strain AS)
FERZA	15	photosystem I, iron–sulfur center protein, rice chloroplast
A23591	15	proteinase inhibitor II, potato
A30893	13	thyroid hormone receptor α-2, human
S04050	12	*matrix (M1) protein, influenza A virus
S00727	10	kinase-related transforming protein (erbB), avian
A27729	11	laminin B2 precursor, mouse
S01358	9	salivary glue protein sgs-3 precursor, fruit fly
A31826	6	*M polyprotein precursor, snowshoe hare bunyavirus
A29438	8	gene frxB protein, wheat chloroplast
A27492	8	hydrogenase small-chain precursor, *Desulfovibrio gigas*
CBUTB	8	cytochrome *b*, *Trypanosoma brucei* mitochondrion
A33095	8	*E2 glycoprotein precursor, mouse hepatitis virus (strain JHM)
OPHUM	8	myeloperoxidase precursor, human
S06546	7	finger protein (clone XlcOF7.1), African clawed frog (fragment)
A33466	7	*basic fibroblast growth factor, African clawed frog
A34373	6	*histidine-rich calcium binding protein precursor, rabbit
A26325	6	*deoxyribonuclease I, bovine (fragment)
A34374	6	*DNA-directed RNA polymerase I, *Trypanosoma brucei*
A25877	6	*cytochrome *c* oxidase polypeptide III, Crithidia
A27878	6	glucuronosyltransferase, human
A34235	5	*complement C6 precursor, human
A25298	5	von Willebrand factor precursor, human
KRGLBS	4	keratin, feather, silver gull
A25468	4	T-cell surface glycoprotein CD3 γ chain
A33896	4	*isotocin–neurophysin I precursor, White sucker
B34141	4	*cytochrome c_{551}, *Pseudomonas aeruginosa*
A32141	4	follistatin 1 precursor, human
HACHPE	4	hemoglobin π′ chain, chicken
A29809	4	tetracycline resistance protein, Campylobacter
S00726	4	kinase-related transforming protein (A-raf), rat
S05984	4	*repB protein, *Bacillus* sp. plasmid pTB913
JL0085	4	*eosinophil granule major basic toxin protein precursor, human
B25103	4	ferredoxin-like protein, *Azotobacter chroococcum*
A27605	4	dystrophin, human
S07508	3	*DNA primase, bacteriophage T3
W6WL	3	probable E6 protein, papillomavirus (type 1a)
P1WL5	3	probable L1 protein, papillomavirus (type 5)
DGECFP	3	formamidopyrimidine DNA glycosylase, *Escherichia coli*
S05342	3	DNA-directed RNA polymerase α chain, Green alga
A33329	3	*testis-specific protein, mouse
QSBPA7	2	host specificity protein A, bacteriophage T7
A20981	3	*transferrin precursor, human
D34443	3	*nitrogen fixation protein nifU, *Anabaena* sp.
A31995	3	nicotinic receptor-associated protein 46K, mouse 2
W6WL31	3	probable E6 protein, papillomavirus (type 31)
S05341	3	probable reverse transcriptase, Green alga KS3/2
A34005	3	*polyergin precursor, Green monkey
A31671	3	H^+/K^+-transporting ATPase, pig
FIMS4A	3	initiation factor eIF-4A, mouse
B35719	3	*phnJ protein plasmid BW120, *Escherichia coli*
QRECLH	3	leucine transport protein livH, *Escherichia coli*
GNMVGV	3	pol polyprotein, AKV murine leukemia virus
B24698	3	formate dehydrogenase B-chain, Methanobacterium
B26635	3	*cytochrome *b* mRNA maturase NAM2 protein 2, yeast (*Saccharomyces cerevisiae*) mitochondrion
WMBEH2	2	UL32 protein, Herpes simplex virus (type 1)
IVRTA1	2	interferon α-I-1 precursor, rat
B24720	2	ORF311 protein, *Bacillus subtilis*
QQBEF2	2	HXLF2 protein, cytomegalovirus (strain AD169)
A32795	2	*T-cell translocation protein 1, human
A35268	2	*nodD protein, *Azorhizobium caulinodans*
A33380	2	interleukin-4 receptor precursor (version 1), mouse
VHWVB	2	structural polyprotein, Sindbis virus (two strains)
S02359	2	29K protein, tobacco rattle virus (strain SYM)
A29270	2	*nifL regulatory protein, *Klebsiella pneumoniae*
A27538	2	complement C5, mouse (fragment)
H35719	2	*phnP protein plasmid BW120, *Escherichia coli*
S01836	2	pyruvate–flavodoxin oxidoreductase, *Klebsiella pneumoniae*
B32352	2	molybdopterin-converting factor chlN, *Escherichia coli*
MNXRS1	2	nonstructural protein NS1, Bluetongue virus
total 556		

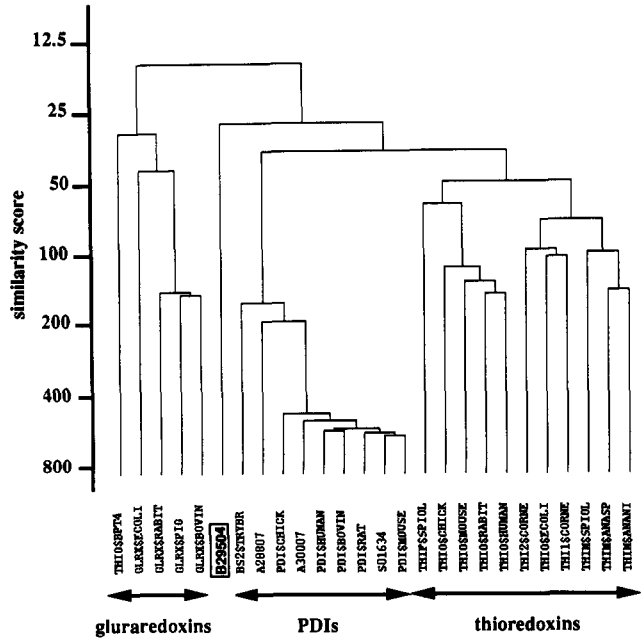

FIGURE 2: Cladogram of the similarity cluster analysis of the functional set and B29504. The higher the similarity score, the more similar are two protein sequences or groups of sequences.

six residues from the active-site CXXC tetrapeptide.

Control Set. The PIR 26 database, containing over 25 000 protein sequences, was searched for all sequences that match pattern 1. From the 670 sequences that match, we removed 40 which either were in the defining functional set or were their homologues. We also removed 47 sequences belonging to hypothetical proteins. The remaining 583 proteins were clustered by similarity score as above. Seventy-seven clusters with 2 or more members were found, containing 556 sequences. The other 27 sequences were singletons, not in any cluster.

One representative from each cluster was used in the control set. Ferredoxins and alcohol dehydrogenases form the two largest clusters. Each of these has one member whose X-ray structure is known, and that member was chosen as the representative. The first member of the other clusters was arbitrarily the representative. As with the functional set, if any representative sequence contained ambiguous residues (X, Z, or B), it could not be used and was replaced by the next in its cluster. PIR locus names and titles of each representative and the number of members in its cluster are given in Table II.

Pattern Development. Pattern 1 contains only a small amount of primary structure information, and small amounts of local protein sequence similarity do not imply a structural

```
(defpattern redox-21
  (pattern
       (b-strand
       (* :gap-min -9 :gap-max -1 :gap-max-overrun 0)
       aliphatic
       (* :gap-min 1 :gap-max 2 :gap-max-overrun 0)
       aromatic
       (* :gap-min 4 :gap-max 6 :gap-max-overrun 0)
       b-turn
       (* :gap-min -5 :gap-max -4 :gap-max-overrun 0)
       (c :score-if-mismatched ,-infinity)
       x x
       (c :score-if-mismatched ,-infinity)
       (* :gap-min -3 :gap-max 0 :gap-max-overrun 0)
       a-helix
       (near-front-of-prev :start-offset 14 :stop-offset 21
       :search-for
       (b-strand) ) )))
```

FIGURE 3: Pattern 3 for the thioredoxin redox motif. The pattern is written in the ARIADNE dialect of Lisp as described in the text.

relationship (Sternberg & Islam, 1990). Indeed, since both the functional and control sets share the same primary sequence similarity (pattern 1), other information is needed to develop a pattern or motif that can distinguish between them.

For this purpose, secondary structure is used, as predicted by a modification of the method of Chou and Fasman (1978) implemented by Ralph et al. (1987). This implementation allows overprediction of secondary structure in order to increase the probability that all true secondary structures will be identified. The related increase of false positive secondary structure predictions is less important in this application.

The predicted secondary structure pattern of the thioredoxin redox motif is

pattern 2

β-strand {0,7} β-turn {−4,0} α-helix (13,20) β-strand

where {x,y} indicates a gap size of from x to y residues of any kind between the end of a predicted secondary structural element and the start of the next element and (x,y) indicates an offset of from x to y residues between the start of one element and the start of the next. A negative gap indicates partial overlap of the following element with the preceding element; thus, "β-turn {−4,0} α-helix" states the helix may begin immediately after the end of the predicted turn or as early as four residues before the end (equivalent to the start of a four-residue β-turn).

Patterns 1 and 2 were initially combined by aligning the CXXC primary sequence on the β-turn predicted secondary structure element. The resulting pattern, containing both primary and secondary structural elements, was then used to search both functional and control sets using ARIADNE, also developed by the MBCRR (Webster et al., 1987). Patterns were iteratively refined using ARIADNE and other tools in the MBCRR Package. The final pattern, pattern 3, is given in Figure 3. Pattern 3 is a combination of the four secondary structural elements in pattern 2 superimposed on the four primary sequence elements in pattern 1 with defined gaps between each element.

Pattern 3 (Figure 3) is formally defined in the ARIADNE dialect of Lisp (Webster et al., 1987). Though written as input to ARIADNE, this pattern is also human-readable. For example, the third through the fifth lines are read as follows: Search for residues that are predicted to be in β-strand secondary structure. When these are found, go to the end of the predicted strand and search for the second pattern element (an aliphatic residue). The gap required between the first and second elements is −9 to −1. A gap size of zero would have the second element directly follow the first. A gap range of −9 to −1 states that the second element is found between the ninth and the first residue from the C-terminal end of the first element.

ARIADNE evaluates how well each sequence matches the pattern. First, if any predicted secondary structural element or specified amino acid class is missing, the match will fail (be reported as a nonmatch). In our pattern, we also require the match to fail if either of the two Cys residues are not present in the proper position. This is accomplished by assigning a score of −infinity if either cysteine is missing (lines 11 and 13 in pattern 3, Figure 3), and avoids spurious matches to patterns that do not include the active-site CXXC.

For each sequence that matches a pattern, ARIADNE computes a score, the sum of partial scores given for each pattern element. Eight pattern elements are specified in pattern 3 (Figure 3): b-strand; aliphatic; aromatic; b-turn; c; c; a-helix; b-strand. Each element is scored in one of four ways: (1) A match with a single amino acid (such as a C) contributes a partial score of 1.0. (2) A match with one member of an amino acid class contributes a partial score of $1/N$ where N is the number of amino acids in the class. For example, the class of aromatic amino acids is F, Y, and W, and a match with one of them would contribute a partial score of 0.33. (3) The partial score for match of a β-turn is 1.0. (4) For β-strands and α-helices, the partial score is the average "strength" of the prediction of that structure for each residue in the predicted strand or helix; in this application, the partial score for strands and helices is between 1.1 and 1.4.

Pattern 3 was developed using the 25 functional and 77 control proteins. All 25 functional proteins and only 3 of the controls matched this pattern. The scores for each of these 28 proteins, computed as described above, are shown in Table III. Table III lists the matches of both control and functional sequences in the training set, and the scores of each match. In analyses of this type, usually only the controls that match with scores higher or equal to a score used as the cutoff are shown. This cutoff score includes the fewest controls above it and the fewest functionals below it. Pattern 3 matches all functionals with scores ≥7.2, a score that includes only two

Table III: Training Set Matches to Pattern 3

set[a]	locus name	score	title
F	THIO$CHICK	7.56	thioredoxin, chicken
F	THIO$RABIT	7.54	thioredoxin, rabbit
F	THIO$MOUSE	7.49	thioredoxin, mouse
F	A30007	7.45	glycosylation site binding protein, chicken
F	PDI$HUMAN	7.42	PDI, human
F	PDI$BOVIN	7.42	PDI, bovine
F	PDI$RAT	7.42	PDI, rat
F	PDI$CHICK	7.42	PDI, chicken
F	GLRX$ECOLI	7.41	glutaredoxin, *E. coli*
F	GLRX$RABIT	7.41	glutaredoxin, rabbit
F	THIM$ANANI	7.39	thioredoxin m, *Anacystic nidulans*
F	THIO$HUMAN	7.39	thioredoxin, human
F	PDI$MOUSE	7.39	PDI, mouse
F	S01634	7.39	thyroid hormone binding protein precursor, mouse
F	BS2$TRYBR	7.38	BS2 protein, *Trypanosoma brucei*
F	THIF$SPIOL	7.37	thioredoxin F, spinach
F	THIM$ANASP	7.37	thioredoxin M, Anabaena
F	GLRX$BOVIN	7.37	glutaredoxin, bovine
F	GLRX$PIG	7.37	glutaredoxin, pig
F	THI1$CORNE	7.35	thioredoxin C-1, *Corynebacterium nephridii*
F	THIO$ECOLI	7.33	thioredoxin, *E. coli*
C	TVVPAS	7.29	large T antigen, polyomavirus BK (strain AS)
F	THIO$BPT4	7.28	thioredoxin, bacteriophage T4
F	A28807	7.28	phosphoinositol–phospholipase C, form I, rat
C	A27492	7.28	hydrogenase small-chain precursor, *Desulfovibrio gigas*
F	THI2$CORNE	7.24	thioredoxin C-2, *Corynebacterium nephridii*
F	THIM$SPIOL	7.20	thioredoxin m, spinach
C	S02359	7.10	29K protein, tobacco rattle virus

[a] F = functional set; C = control set (see text).

Table IV: Two-by-Two Table of Redox Pattern 3 for the Training Set[a]

	HIT >= 7.1966662	MISS < 7.1966662	
FCNL	25	0	25
CNTL	2	75	77
	27	75	102

[a] Sensitivity = 1.0; specificity = 0.974; 2 × 2 correlation = 0.9497.

controls. We have included for information the only other matching control; it is a true negative and is listed in Table III in the bottom row, separating it from the true positive (functional) and false positive (control) matches.

A summary of the matches in a two-by-two table is produced by ARIADNE, as shown in Table IV. In Table IV, a "hit," has a score ≥7.2 and a "miss" has a score of <7.2. All of the functional sequences (FCNL) match (hit) the pattern with scores ≥7.2, while only 2 of the 77 control sequences (CNTL) match. This results in a very high sensitivity and specificity. Sensitivity, or true-positive ratio, is the likelihood that a protein with the structure matches the pattern; it equals the number of true positives divided by the sum of true positives and false negatives. Specificity, or true-negative ratio, is the likelihood that a protein lacking the structure does not match the pattern; it equals the number of true negatives divided by the sum of true positives and false negatives. As shown in Table IV, pattern 3 has a sensitivity of 1.0 and a specificity of 0.97.

Control Sequences Not in the Training Set. As described above, control sequences are those that match pattern 1, the initial screening pattern, and have no reported thioredoxin-like functionality. Control sequences have a hydrophobic aliphatic residue (V, I, or L) separated by one or two residues from an aromatic residue (F, Y, or W) separated by three to six residues from the active-site CXXC tetrapeptide. Since only 77 of the 583 control sequences were used in the training set (as described above), to further characterize false-positive matches the entire group of 583 was studied. Since some control proteins are very large and we are interested in only a small domain, all were truncated ≤100 amino acids on either side of the sequence that matches pattern 1. Four sequences were excluded because they contain ambiguous residues. From 579 control sequences, 587 nonoverlapping segments containing pattern 1 were matched against pattern 3. Thus, pattern 3 has been tested against *all* proteins in the 25 000-sequence PIR that are potential matches. The 13 control proteins that match with scores equal to or greater than the lowest functional match and the 6 control proteins that match with lower (nonfunctional) scores are shown in Table V.

Control Matches. Of the 13 control matches listed in Table V, 2 are essentially duplicates (large T antigen), reducing the number to 12. We consider these the only false positives in over 25 000 PIR sequences. Several lines of evidence support the view that these are false positives. While there are no proteins in Table V with known X-ray structures, there are crystal structures of two proteins related to a number of them. These are FESGAL (ferredoxin from *Spirulina platensis*) and DEHOAL (alcohol dehydrogenase from horse liver). FESGAL and DEHOAL match the pattern 1 primary sequence in cysteine-rich regions involved with the iron–sulfur or zinc–sulfur clusters, respectively. FESGAL and DEHOAL do not match pattern 3 nor do they have thioredoxin functionality in the regions that match pattern 1. Similar involvement with metal–sulfur clusters is suggested for the primary sequence segments that match pattern 3 in the ferredoxin, polyferredoxin, alcohol dehydrogenase, and hydrogenase sequences. These ferridoxin- and ADH-related sequences account for 7 of the 12 control matches.

Table V: Control Proteins That Match Pattern 3

PIR name	score	title
S02006	7.51	phosphoprotein phosphatase type X catalytic chain, rabbit (fragment)
NJBYM1	7.36	M1-1 protoxin precursor, yeast
A32315	7.30	*hydrogenase [NiFe] small-chain precursor, *Desulfovibrio gigas*
TVVPAS	7.29	large T antigen, polyomavirus BK (strain AS)
TVVPTB	7.29	large T antigen, polyomavirus BK
A27492	7.28	hydrogenase small-chain precursor, *Desulfovibrio gigas*
S06200	7.28	alcohol dehydrogenase 1, White clover
JV0072	7.26	hydrogenase small chain, *Escherichia coli*
G30315	7.25	*polyferredoxin, *Methanobacterium*
A27689	7.24	complement C5, human (fragment)
S00912	7.23	alcohol dehydrogenase 1 (clone lambda-PG8), garden pea
A25691	7.21	glucocorticoid receptor, mouse
S08393	7.21	*ferredoxin, plant-type, *Rhodobacter capsulatus*
	7.20	**lowest functional match (see Table IV)**
A31341	7.19	hydrogenase small-chain precursor, *Bradyrhizobium*
S07391	7.17	*gene algD protein, *Pseudomonas aeruginosa* (GDP-mannose dehydrogenase)
C8HUA	7.11	complement component C8 α-chain precursor, human
WMBV2P	7.10	28.8K protein, tobacco rattle virus (strain PSG)
S02359	7.10	29K protein, tobacco rattle virus
A27538	7.07	complement C5, mouse (fragment)

Several control matches are the only matches in their clusters. These "single matches" account for an additional 3 of the 12 control matches. Since the other members of the cluster do not match, a single match is most likely fortuitous and with no structural significance for the cluster as a whole. S02006, the highest scoring control, is a phosphoprotein phosphatase, and the only one of 27 phosphoprotein phosphatases in the control sequences to match. Similarly, A25691, the mouse glucocorticoid receptor, is the only one of the 16 sequences in the hormone receptor cluster to match. A25691 has 3 CXXC's and 3 additional C's within 30 residues. As mentioned above, mammalian glucocorticoid receptors are activated by thioredoxin (Grippo et al., 1985). Also TVVPAS and TVVPTB, the nearly identical viral large T antigens, are the only ones in the nine-member T antigen cluster which match.

The remaining 2 of the 12 control matches are also considered false positives. A27689 (as well as the A27538 below the cutoff), the large fragment of complement C5, matches pattern 3 at its C-terminal end. For A27689, pattern 3 ends at residue 1284 out of 1284 and for A27538 at residue 1668 out of 1671. There are no residues for the remaining α_3–β_4–β_5–α_4 structure of a thioredoxin domain, and we consider this a spurious match. For NJBYM1, the yeast protoxin precursor, the two C's in the CXXC region are not bound to each other. Rather, each is involved in a structurally important disulfide bond with a C in a linearly distant part of the sequence.

Gonadotropic Hormones. Boniface and Reichert (1990) reported "thioredoxin-like" redox activity for follotropin and lutropin; thus, it was of interest to test our pattern on this class of proteins. The PIR and Swiss-Prot databases were searched, and five follotropin and five lutropin sequences were extracted. Loci names and sequence titles are given in Table VI. While more sequences were found, these were duplicates or precursors

Table VI: Gonadotropic Hormone Set

PIR or Swiss-Prot locus name	title
LSHB$SHEEP	lutropin, sheep
LSHB$BOVIN	lutropin, bovine
LSHB$HUMAN	lutropin, human
LSHB$HORSE	lutropin, horse
LSHB$RAT	lutropin, rat
FSHB$SHEEP	follotropin, sheep
FSHB$BOVIN	follotropin, bovine
FSHB$HUMAN	follotropin, human
FTHOB	follotropin, horse
A32893	follotropin, rat

Table VII: Hypothetical Proteins That Match Pattern 3

PIR locus name	score	title
B29504	7.38	*mer* operon 18K hypothetical protein, *Staphylococcus aureus* plasmid pI258
S03229	7.29	hypothetical protein B-129, *Sulfolobus acidocaldarius* virus-like particle SSV1
I30010	7.27	hypothetical HURF-5 protein, *Sauroleishmania tarentolae* mitochondrion (SGC6)
	7.20	**lowest functional match (see Table IV)**
QQBE6	7.07	hypothetical BFLF1 protein, Epstein–Barr virus (strain B95–8)

of those included. None of these 10 sequences match pattern 3; only A32893 (follotropin from rat) matches the primary structural elements (pattern 1), and all lack a predicted α-helix secondary structural element immediately following the CXXC.

PDIs and PDI Analogues. Pattern 3 matches only the second (C-terminal) of the two thioredoxin domains in PDI and PDI analogues. In an attempt to match the first one, the maximum value of the last gap (that between the α-helix and the trailing β-strand) was increased by changing the stop-offset parameter from 21 to 24. This is equivalent to permitting an additional three residues between the predicted start of the helix and the start of the strand. This relaxed pattern 3 will match the first thioredoxin-like domain in all PDIs and the closely analogous mouse thyroid hormone binding protein in Table I. The relaxed pattern matches one additional member of the control set (A29270, score = 7.29), so the sensitivity remains 1.0 but the specificity is reduced from 0.97 to 0.96. The relaxed pattern does not match the first thioredoxin-like domain in the less similar PDI analogues: chicken glycosylation site binding protein, rat phosphoinositol–phospholipase C, and the protein with unknown function from *Trypanosoma brucei*.

The protein coded for by the *dsbA* mutation in *E. coli* (Bardwell et al., 1991), just as the N-terminal thioredoxin domain in PDI is not matched by the original pattern 3, but is matched by the relaxed pattern.

Hypothetical Proteins. We next tested pattern 3 against the 47 hypothetical proteins removed from the initial control sequences. Table VII lists the three hypothetical sequences that match pattern 3 with a score >7.2 and the only other match (score 7.1). The highest scoring hypothetical sequence, B29504, is ORF3 from the *mer* operon of *Staph. aureus*.

Functional Set Clustering. A tree diagram of the similarity clustering of the functional set is shown in Figure 2. The tree displays the similarity score in a cladogram on a log scale with higher (more similar) scores lower on the ordinate. The further down in Figure 2 two sequences are connected, the more similar they are. For example, all PDIs are very similar (are connected near the bottom of the figure), with scores >500.

B29504, the highest scoring hypothetical sequence, is included in the functional set similarity clustering in Figure 2. B29504 is weakly clustered (score 28) with the thioredoxins and PDIs. However, as shown in Figure 2, the similarity clustering score is higher (stronger) between B29504 and the thioredoxin–PDI cluster than between the thioredoxin–PDI cluster and the glutaredoxin cluster (score 16).

Alignments and Hydrophobicity Patterns. The sequence alignments for the amino acid sequences that match pattern 3 and surrounding residues for *E. coli*, T4, and *Corynebacterium nephridii* thioredoxins and *E. coli* glutaredoxin, *E. coli* DsbA protein, and the three matching hypotheticals (B29504, S03229, and I30010) are shown in Figure 4. The α-helix and β-strand secondary structure assignments are based on X-ray crystallography for *E. coli* and T4 thioredoxins (Holmgren et al., 1975; Katti et al., 1990; Söderberg et al., 1978) and from the sequence alignment for *E. coli* glutaredoxin and *Coryneform bacterium* thioredoxin (Eklund et al., 1984).

The thioredoxins and glutaredoxins were aligned on the basis of Eklund et al. (1984), and the hypothetical proteins S03229 and I30010 were aligned on the CXXC in pattern 3. B29504 was aligned with the thioredoxins and glutaredoxins by first eliminating the 24-residue, N-terminal, hydrophobic probable signal sequence (Laddaga et al., 1987). Next PIMA was used to align this sequence with *E. coli* thioredoxin (the thioredoxin with known X-ray structure). DsbA was aligned in a similar manner. The first 17 residues were removed; then it was manually aligned on the CXXC.

Finally, the B29504 and DsbA alignments were manually modified to optimize the fit to hydrophobicity patterns of the thioredoxins and glutaredoxins, align the conserved proline residue at the N-terminus of strand β_4, and place gaps between rather than within areas of predicted secondary structure. Hydrophobicity patterns are shown in stylized format between the sequences in Figure 4. The C and P residues used in the alignment are retained, and hydrophobic residues (F, W, Y, V, I, M, C, and L) are symbolized as (■).

The stylized format for display of hydrophobicity patterns simplifies visual comparisons. Manual alignment using this format permits prediction of secondary structure similarity for regions outside pattern 3 in both B29504 and DsbA, but not for S03229 or I30010.

Discussion

Quality of the Pattern. An ARIADNE pattern is a description of a unit of tertiary structure defined in terms of its linear array of secondary structural elements, variable sized gaps, and a few amino acid residues. This is illustrated in Figure 5 where the unit of thioredoxin tertiary structure (top) is linearized to more clearly show the sequential arrangement of secondary structure elements and gaps. The essential thioredoxin pattern is the following: strand–gap–CXXC–helix–gap–strand. Pattern 3 includes this and, in addition, specifies two amino acids in the first strand. In practice, a pattern is developed, and its parameters are optimized iteratively to include functional sequences and exclude control sequences.

The quality of a pattern or descriptor is evaluated by sensitivity and specificity, which have maximum values of 1. These performance indexes are very high for the descriptor reported here; pattern 3 has a sensitivity of 1.0 and a specificity of 0.97 (Table IV). For comparison, a good ARIADNE pattern for transcriptional activation among nuclear and DNA binding prokaryotic proteins is reported to have a sensitivity of 0.7 and a specificity of 0.9 (Zhu et al., 1990).

The sensitivity and specificity indexes are computed from data of the type in Table III. Examination of the matches listed in Table III illustrates the high performance of pattern 3. For a cutoff score of 7.2, all functional sequences and only

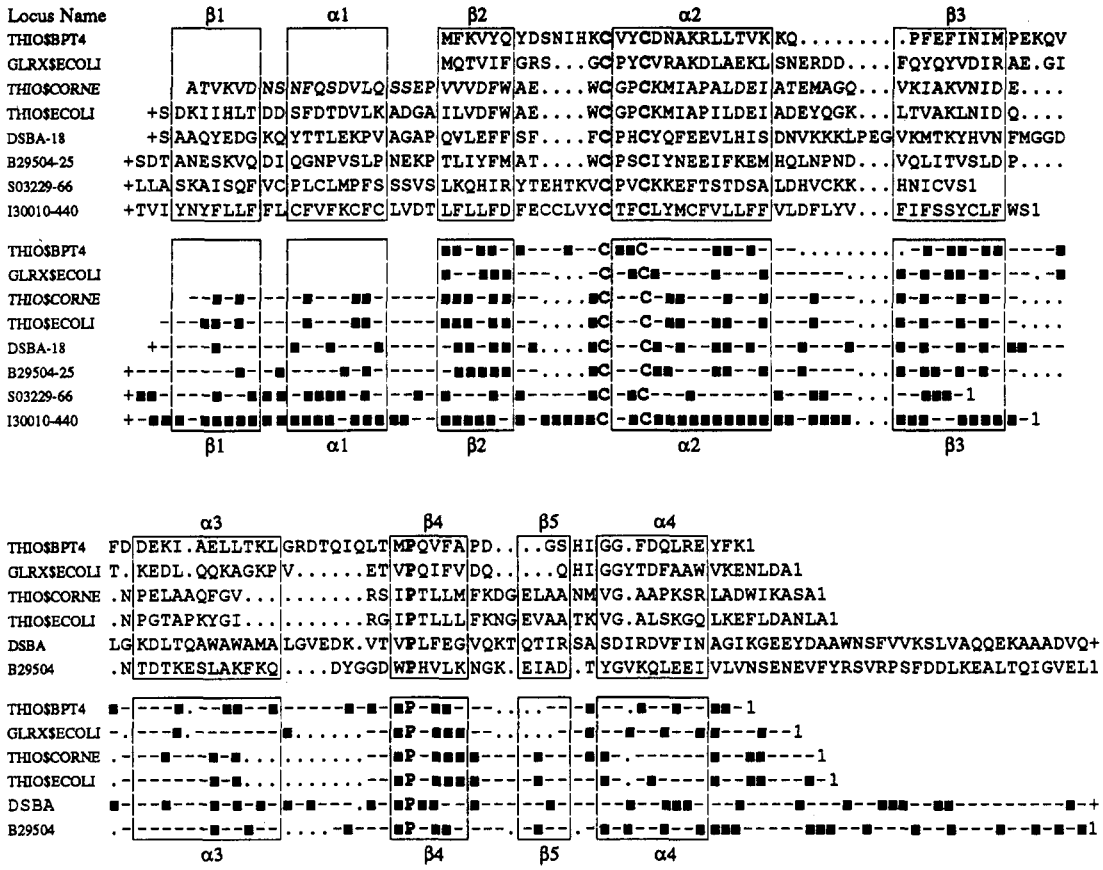

FIGURE 4: Sequence alignments of thioredoxins and glutaredoxins with hypothetical proteins that match pattern 3. The secondary structures of THIO$ECOLI (*E. coli* thioredoxin) and THIO$BPT4 (T4 thioredoxin) are taken from their crystal structures. Sequence alignment is described in the text. Abbreviations: C and P, cysteine and proline residues, respectively, used in the alignment; (■) hydrophobic residues (F, W, Y, V, I, M, C, L); (-) all other residues; (.) gap; 1, C-terminus; (+) additional N- or C-terminal residues. A number after the hyphen in the locus name indicates the ordinal number of the first residue in the alignment. The 17-residue leader sequence not shown for DSBA is MKKIWLALAGLVLAFSA-. The 43-residue terminal sequence not shown for DSBA is -LRGVPAMFVNGKYQLNPQGMDTSNMDVFVQQYADTVKYLSEKK1. The 24-residue leader sequence not shown for B29504 is MKKRISFTAIMTVLLIGLTACGAE-.

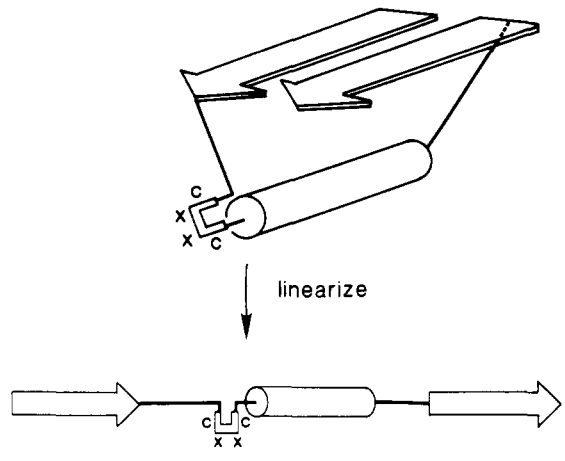

FIGURE 5: Diagram of the tertiary structural motif of the active site of thioredoxins. Pattern 3 (Figure 3) is a description of the linear arrangement (bottom) of secondary structural elements [after Webster et al. (1987)].

two control sequences match pattern 3. That is, a cutoff of 7.2 includes a very small overlap of control sequences while including all the functional sequences.

For a pattern of this type, restricted to a relatively small number of proteins (25) from 1 evolutionary family, the specificities and sensitivities calculated here are upper bounds. Ideally, we would like to test the pattern against a comparable set of known functionals and controls not used in developing the pattern. However, this has not been done here, nor by others in the development of similar patterns. With only 25 members in the functional set, any group withheld as members of a testing set would not constitute an equivalent but independent set of proteins. When more complete genome sequences become available, it might be possible to answer questions of this type by testing patterns against all reading frames in a genome (or a representative portion of it) and only then deriving "true" sensitivity and specificity.

Gonadotropic Hormones. Thioredoxin-like redox functionality has been proposed for lutropin and follitropin, in part on the basis of their having CXPC sequences (Boniface & Reichert, 1990). While one (A32893) does match the primary sequence elements of our pattern, all tested lack the predicted α-helix immediately following the CXXC (Table VI). These hormones may have physiologic redox functionality; our results suggest it does not reside in a thioredoxin structural domain.

PDI and PDI Analogues. Pattern 3 matches only the C-terminal thioredoxin-like domain in PDIs and their analogues. However, if this pattern is relaxed (altered) slightly to increase the number of residues in the gap between the α-helix and the trailing β-strand, the N-terminal thioredoxin-like domain is matched in PDIs, but not in most PDI analogues.

A result similar to that found for the N-terminal PDI domain is observed for the DsbA protein. This protein has PDI redox functionality and contains a Cys-Pro-His-Cys region that resembles the thioredoxin and PDI active sites. The authors suggest that it is either a variant PDI or a member of a new

family of oxidoreductases (Bardwell et al., 1991). This protein does match the relaxed pattern 3, but not pattern 3, implying it has a more extended loop between α_2 and β_3. This is interesting because the α_2–β_3 loop is on the opposite side of the molecule from the active site (arrow in Figure 1), and may be related to binding of ligands other than the redox substrate.

The hydrophobicity patterns of thioredoxin-related proteins, described under Results and shown in Figure 4, support assignment of DsbA to the thioredoxin superfamily. Alignment of the CXXC in DsbA and thioredoxins, and inclusion of an extended loop between α_2 and β_3 in DsbA, gives highly similar hydrophobicity patterns in the predicted β_2, α_2, β_3, α_3, β_4, and β_5 secondary structural elements (Figure 4).

Hypothetical Proteins. Three of the 47 hypothetical proteins that match pattern 1 also match pattern 3 above the cutoff (Table VII). The alignments for the portions of the sequences that match pattern 3 for the three matching hypotheticals (B29504, S03229, and I30010) and *E. coli*, T4, and *Corynebacterium nephridii* thioredoxins and *E. coli* glutaredoxin are shown in Figure 4. Two of the three hypothetical matches are considered false positives. As found for two of the false control matches, these sequences (S03229 and I30010) match pattern 3 at their C-termini (Figure 4). Pattern 3 ends at the terminal residue 129 in S03229, and at the penultimate residue 501 in I30010. While it is possible that the β–α–β unit of structure is present in these proteins, it cannot be incorporated into a full thioredoxin domain. Additionally, the hydrophobicity patterns for these sequences show little similarity to those of thioredoxins and glutaredoxins (Figure 4). I30010 is very hydrophobic throughout and not likely to be a soluble protein.

The most interesting match of the three hypotheticals is B29504, a 161-residue protein from ORF3 of the *mer* operon in *Staphylococcus aureus*. Three considerations lead us to conclude that this protein is likely to have a structure very similar to that of thioredoxin. First, the hydrophobicity pattern of B29504 (Figure 4) is remarkably similar to thioredoxins and glutaredoxins. In addition to the β–α–β unit of pattern 3, this similarity includes the β_1, α_3, β_4, β_5 and α_4 secondary structural elements which are not part of pattern 3. Even without its 24-residue signal sequence (discussed above, and given in the legend of Figure 4), B29504 is longer than most thioredoxins. Assuming that B29504 has a structure similar to thioredoxin, the distribution of hydrophobic residues past α_4 suggests the existence of two additional secondary structure elements at its C-terminus.

Second, the similarity clustering score shown in Figure 2 is higher (stronger) between B29504 and the thioredoxin–PDI cluster than between the thioredoxin–PDI and glutaredoxin clusters. The significance of the clustering pattern in Figure 2 is that it distinguishes between proteins as similar in structure and function as the glutaredoxins, thioredoxins, and PDIs. The first branch of the cladogram separates glutaredoxins from the thioredoxin–PDI cluster and B29504. That is, B29504 is more similar to thioredoxin–PDIs than are glutaredoxins.

Third, the *mer* operon of *Staphylococcus aureus*, which contains the gene for B29504, also codes for a reductase similar to *E. coli* thioredoxin reductase, making it plausible that a thioredoxin-related protein has coevolved with this operon in the Gram-positive bacteria. The *mer* operon contains seven ORFs, two of which have been identified as the *mer A* (mercuric reductase) and *mer B* (organomercural lyase) genes. The reductase detoxifies mercury by reducing Hg^{2+} and Hg^{+} ions to elemental Hg^0. The lyase cleaves the carbon–mercury bond of organomercurials such as phenylmercuric acetate. One product is Hg^{2+}, which is subsequently detoxified by the mercuric reductase [see Laddaga et al. (1987) and references cited therein]. Since it contains a canonical leader sequence, ORF3 is suggested to be an exported binding protein functioning in this pathway (Silver & Misra, 1988). There is no match to a thioredoxin-related motif (pattern 3) in the proteins encoded by the Gram-negative bacterial *mer* operon.

The high sensitivity and specificity of pattern 3 might lead to an expectation of more than one interesting hypothetical match out of the three. However, the pattern was developed on a population of proteins with known function. Hypothetical proteins form a different population. ORFs and their putative protein sequences are typically analyzed by methods [for example, Doolittle (1987)] which identify protein classes and functions from strong similarities to known proteins. Two other ORFs in the *mer* operon were assigned function in this manner (Laddaga et al., 1987). Those ORFs labeled "hypothetical" are without strong similarity to *any* existing protein with known function. The population of hypothetical proteins has already been culled for matches to standard sequences, and by this, it is biased in unknown directions (e.g., membrane associated, structural, not necessarily translated). Thus, even using a pattern with high sensitivity and specificity, there may be more false matches in a population of hypothetical than in a population of known proteins, and all matches to patterns by hypothetical proteins must be analyzed further.

Applicability of ARIADNE *to Other Systems.* The usefulness of ARIADNE in producing a description of the redox motif of thioredoxins depends primarily on the existence of suitable training sets. The transferability of this approach to other proteins is contingent on the availability of a training set with similar characteristics. First, sequences from diverse phyla are important. As few as 10 sequences spanning the diversity of the 25 available for thioredoxin (Figure 2) could be sufficient. If, however, the 10 were closely related, such as our vertebrate PDI sequences, it would be difficult to determine valid pattern elements. Second, it is very helpful to know at least one representative three-dimensional structure to guide initial pattern design. Secondary structure prediction is still rudimentary. Our confidence in the validity of the method we use is strengthened by the accuracy of its prediction of the known secondary structure of *E. coli* and T4 thioredoxins near the active site. Finally, known activity for the functional group, and known absence of this activity in the control group, is important. Without this, it is difficult to evaluate pattern matches in the control set or mismatches in the functional set.

The limitation of the ARIADNE method of pattern development is that it is best used for short, contiguous peptide sequences. Each secondary structural pattern element has an unspecified length, and each is separated from the next by a gap of specified minimum and maximum length. Because there is an uncertainty in the position of any element, the discriminatory capability of the pattern decreases with an increasing number of elements, and there is a limit to the number of structural elements which can sensibly be strung together. In future work, we plan to develop methods for the inclusion of structural elements that are distant in the primary sequence, but close to the active site in the tertiary structure.

CONCLUSIONS

We have generated a pattern descriptor for the tertiary structural motif containing the thioredoxin active site with both high specificity and high sensitivity. A search of the hypothetical protein sequences produces a match to protein B29504, a putative gene product of ORF3 in the *mer* operon of *Staphylococcus aureus*. Secondary structure alignment of B29504 to thioredoxins produces a hydrophobicity pattern very similar

to thioredoxins, even in secondary structural elements not included in the pattern descriptor. Similarity cluster analysis groups B29504 more closely to thioredoxins than thioredoxins are clustered to glutaredoxins. From this and considerations of the *mer* operon, we conclude that B29504 is highly likely to have a structure quite similar to thioredoxin. Our thioredoxin redox pattern suggests that the second thioredoxin-like domain in PDIs is more similar to thioredoxin than the first and that the first thioredoxin-like domain in PDIs is more similar to the DsbA protein than the second. The absence of a match to gonadotropic hormones also suggests that these do not have a thioredoxin domain. The pattern is suitable for use in screening new sequences for thioredoxin domains.

ACKNOWLEDGMENTS

We thank Temple Smith, Teresa Webster, Richard Lathrop, Florence Gleason, and Rebecca Chapin for helpful discussions and advice.

REFERENCES

Bardwell, J. C. A., McGovern, K., & Beckwith, J. (1991) *Cell 67*, 581–589.

Barton, G. J., & Sternberg, M. J. E. (1990) *J. Mol. Biol. 212*, 389–402.

Bashford, D., Chothia, C., & Lesk, A. M. (1987) *J. Mol. Biol. 196*, 199–216.

Benner, S. A., & Gerloff, D. (1990) *Adv. Enzyme Regul. 31*, 121–181.

Boado, R. J., Campbell, D. A., & Copra, I. J. (1988) *Biochem. Biophys. Res. Commun. 155*, 1297–1304.

Boniface, J. J., & Reichert, L. E. (1990) *Science 247*, 61–64.

Bowie, J., Luthy, R., & Eisenberg, D. (1991) *Science 253*, 164–170.

Cheng, S., Gong, Q., Parkison, C., Robinson, E. A., Appella, E., Merlino, G. T., & Pastan, I. (1987) *J. Biol. Chem. 262*, 11231–11277.

Chou, P. Y., & Fasman, G. D. (1978) *Annu. Rev. Biochem. 47*, 251–276.

Cohen, F. E., Arbanel, R. M., Kuntz, I. D., & Fletterick, R. J. (1986) *Biochemistry 25*, 266–275.

Das, A., Hummel, B., Gleason, F., Holmgren, A., & Walfish, P. (1988) *Biochem. Cell. Biol. 66*, 460–464.

Davison, D., & Chapplear, J. E. (1990) *Nucleic Acids Res. 18(6)*, 1571–1581.

Doolittle, R. E. (1987) *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*, University Science Books, Mill Valley, CA.

Edwards, M. S., Sternberg, M. J. E., & Thornton, J. M. (1987) *Protein Eng. 1*, 173–181.

Eklund, H., Cambillau, C., Sjöberg, B.-M., Holmgren, A., Jörnvall, H., Höög, J.-O., & Brändén, C.-I. (1984) *EMBO J. 3*, 1443–1449.

Eklund, H., Gleason, F. K., & Holmgren, A. (1991) *Proteins: Struct., Funct., Genet. 11*, 13–28.

Fasman, G. D. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., Ed.) pp 193–316, Plenum Press, New York.

Faulkner, D., & Jurka, J. (1988) *Trends Biochem. Sci. (Pers. Ed.) 13*, 321–322.

Finkelstein, A. V., & Reva, B. (1991) *Nature 351*, 497–499.

Freedman, R. B., Bulleid, N. J., Hawkins, H. C., & Paver, J. L. (1989) *Biochem. Soc. Symp. 55*, 167–192.

Fuchs, J. A. (1989) in *Glutaredoxin and Glutathione: Chemical, Biochemical and Medical Aspects, Part B* (Dolphin, D., Polson, R., & Avramovie, O., Eds.) pp 551–570, John Wiley & Sones, Inc., New York.

Geetha-Habib, M., Holva, R., Kaplan, H. A., & Lennarz, W. J. (1988) *Cell 54*, 1054–1060.

Gleason, F., & Holmgren, A. (1988) *FEMS Microbiol. Rev. 54*, 271–298.

Grippo, J., Holmgren, A., & Pratt, W. (1985) *J. Biol. Chem. 260*, 93–97.

Holmgren, A. (1985) *Methods Enzymol. 113*, 525–540.

Holmgren, A. (1989) *J. Biol. Chem. 264*, 13963–13966.

Holmgren, A., Söderberg, B.-O., Eklund, H., & Brändén, C.-I. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 2305–2309.

Hsu, M., Muhich, M. L., & Boothroyd, J. C. (1989) *Biochemistry 28*, 6440–6446.

Hunt, T., Herbert, P., Campbell, E., Delidakis, C., & Jackson, R. (1983) *Eur. J. Biochem. 131*, 303–311.

Katti, S., LeMasters, D., & Eklund, H. (1990) *J. Mol. Biol. 212*, 167–184.

Langsetmo, K., Fuchs, J., & Woodward, C. (1989) *Biochemistry 28*, 3211–3222.

Langsetmo, K., Fuchs, J., & Woodward, C. (1991) *Biochemistry 30*, 7609–7614.

Lathrop, R. H., Webster, T. A., & Smith, T. F. (1987) *Commun. ACM 30*, 909–921.

Lim, V. I. (1974) *J. Mol. Biol. 88*, 857–872.

Maness, P., & Orengo, A. (1975) *Biochemistry 14*, 1484–89.

Pihlajaniemi, T., Helaakoski, T., Tasanen, K., Myllyla, R., Huhtala, M. L., Koivu, J., & Kivirikko, K. I. (1987) *EMBO J 6*, 643–649.

Ralph, W. W., Webster, T. A., & Smith, T. F. (1987) *Comput. Appl. Biosci. 3*, 211–216.

Schürmann, P., Maeda, K., & Tsugita, A. (1981) *Eur. J. Biochem. 116*, 37–45.

Silver, S., & Misra, T. (1988) *Annu. Rev. Microbiol. 42*, 717–743.

Smith, R. F., & Smith, T. F. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 118–122.

Smith, R. F., & Smith, T. F. (1992) *Protein Eng. 5*, 35–41.

Söderberg, B.-O., Sjöberg, B.-M., Sonnerstam, U., & Brändén, C.-I. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 5827–5830.

Sternberg, J. M., & Islam, S. A. (1990) *Protein Eng. 4*, 125–131.

Taylor, W. R. (1986) in *Computer Graphics and Molecular Modeling* (Fletterick, R., & Zoller, M., Eds.) pp 77–84, Cold Spring Harbor Laboratory Press, Cold Springer Harbor Laboratory, NY.

Taylor, W. R. (1988) *Protein Eng. 2*, 77–86.

Thornton, J. M., Flores, T. P., Jones, D. T., & Swindells, M. B. (1991) *Nature 354*, 105–106.

Webster, T. A., Lathrop, R. H., & Smith, T. F. (1987) *Biochemistry 26*, 6950–6957.

Wolosiuk, R., & Buchanan, B. (1977) *FEBS Lett. 81*, 253–258.

Zhu, Q., Smith, T. F., Lathrop, R. H., & Figge, J. (1990) *Proteins: Struct., Funct., Genet. 8*, 156–163.