Article No. mb981993 *J. Mol. Biol.* (1998) **281**, 949–968

JMB

Method for Prediction of Protein Function from Sequence using the Sequence-to-Structure-to-Function Paradigm with Application to Glutaredoxins/Thioredoxins and T_1 Ribonucleases

Jacquelyn S. Fetrow[1] and Jeffrey Skolnick[2]*

[1]*Department of Biological Sciences, Center for Biochemistry and Biophysics University at Albany, SUNY 1400 Washington Avenue Albany, NY 12222, USA*

[2]*Department of Molecular Biology, The Scripps Institute 10550 North Torrey Pines Road, La Jolla, CA 92037, USA*

**Corresponding author*

The practical exploitation of the vast numbers of sequences in the genome sequence databases is crucially dependent on the ability to identify the function of each sequence. Unfortunately, current methods, including global sequence alignment and local sequence motif identification, are limited by the extent of sequence similarity between sequences of unknown and known function; these methods increasingly fail as the sequence identity diverges into and beyond the twilight zone of sequence identity. To address this problem, a novel method for identification of protein function based directly on the sequence-to-structure-to-function paradigm is described. Descriptors of protein active sites, termed ''fuzzy functional forms'' or FFFs, are created based on the geometry and conformation of the active site. By way of illustration, the active sites responsible for the disulfide oxidoreductase activity of the glutaredoxin/thioredoxin family and the RNA hydrolytic activity of the T_1 ribonuclease family are presented. First, the FFFs are shown to correctly identify their corresponding active sites in a library of exact protein models produced by crystallography or NMR spectroscopy, most of which lack the specified activity. Next, these FFFs are used to screen for active sites in low-to-moderate resolution models produced by *ab initio* folding or threading prediction algorithms. Again, the FFFs can specifically identify the functional sites of these proteins from their predicted structures. The results demonstrate that low-to-moderate resolution models as produced by state-of-the-art tertiary structure prediction algorithms are sufficient to identify protein active sites. Prediction of a novel function for the gamma subunit of a yeast glycosyl transferase and prediction of the function of two hypothetical yeast proteins whose models were produced *via* threading are presented. This work suggests a means for the large-scale functional screening of genomic sequence databases based on the prediction of structure from sequence, then on the identification of functional active sites in the predicted structure.

Keywords: protein function prediction; *ab initio* folding algorithm; threading algorithm; ribotoxin; functional genomics

Introduction

The Human Genome Project began with the specific goal of obtaining the complete sequence of the human genome and determining the biochemical nature of each gene. To date, the project has been quite successful, with sequencing of the human

Present addresses: J. S. Fetrow, Department of Molecular Biology, The Scripps Institute, LaJolla, CA 92037, USA

Abbreviations used: ORF, open reading frame; FFF, fuzzy functional forms.

Email address of the corresponding author: skolnick@scripps.edu

0022–2836/98/350949–20 $30.00/0

genome about 1.2% complete (J. Roach, http://weber.u.washington.edu/~roach/human_genome_progress2.html; Gibbs, 1995), and is on track for its scheduled completion in the year 2005. Furthermore, the genomes of 14 organisms have been sequenced and published, including *Mycoplasma genitalium* (Fraser *et al*., 1995), *Methanococcus jannaschii* (Bult *et al*., 1996), *Haemophilus influenzae* (Fleischmann *et al*., 1995), *Escherichia coli* (Blattner *et al*., 1997) and *Saccharomyces cerevisiae* Mewes *et al*., 1997). Significant progress has been made in mapping and sequencing the genomes of model eukaryotic organisms, such as mouse, *Caenorhabditis elegans* and *Drosophila melanogaster*.

One of the goals of the genome project is to develop tools for comparing and interpreting the resulting genomic information (Collins & Galas, 1993). Researchers must learn where each gene lies and must understand the function of each gene or gene product: is the nucleotide sequence a regulatory region? Does the nucleotide segment produce a gene product? Is the product active as an RNA or a protein molecule? What function does the gene product perform: does it bind to another molecule, is it important for regulation of cellular processes, does it catalyze a chemical reaction? The importance of answering these questions has led to research efforts directed towards understanding or describing the function of each sequence, particularly for protein sequences and open reading frames (ORFs). Most often functional analysis is done by sequence comparison to proteins of known structure or function; however, because of the lack of sequence similarity, these methods fail on about half of the sequences available in the sequence and genome databases (Delseny *et al*., 1997; Dujon, 1996). Other approaches to function prediction include comparison of the complete (Himmelreich *et al*., 1997) microbial genomes sequenced thus far and an analysis of gene clustering (Himmelreich *et al*., 1997: Tamames *et al*., 1997). Some have proposed experimental methods to accomplish aspects of function prediction on a genome-wide basis (Fromont-Racine *et al*., 1997; Sakaki, 1996). Here, in contrast, we present a novel method for protein function prediction based on the sequence-to-structure-to-function paradigm, where the protein structure is first predicted from the sequence, then the active site is identified within the predicted structure. Thus, this method requires only knowledge of the protein primary sequence. As will be demonstrated, enzyme active sites can be specifically identified in structures produced by state-of-the-art prediction algorithms where the atomic coordinates are not well defined.

Sequence alignment methods for function identification

The most common method of function identification from just the sequence is global or local sequence alignment. This technique is based on finding the extent of sequence identity between a given sequence and another whose function is known. Significant sequence identity is a strong indicator that the proteins probably have similar functions. Alignment methods such as BLAST (Altschul *et al*., 1990), BLITZ (MPsrch; Sturrock & Collins, 1993), and FASTA (Pearson & Lipman, 1988), among others, are currently the most powerful techniques for analyzing the many sequences found in the genome databases. Today's methods are robust, fast and powerful for determining the relatedness of protein sequences, particularly when the sequence identity is above 30% and the relationship between proteins is unequivocal.

Limits to sequence alignment methods

A major problem with sequence alignment methods for analysis of protein function arises when the sequence similarity goes below the twilight zone of 25 to 30% sequence identity. Currently available programs cannot consistently detect functional and structural similarities when the sequence identity is less than 25% (Hobohm & Sander, 1995). Matches with 50% amino acid identity over a 40 residue or shorter stretch of sequence regularly occur by chance and relationships between such proteins must be viewed with caution, unless other information is available (Pearson, 1996). In the worst case, protein sequences or ORFs do not return significant matches to any sequences in the database. For instance, experiments showed that an ORF from an intron in a cyanobacterium tRNA (Biniszkiewicz *et al*., 1994) was found to produce a protein with endonuclease activity, but no significant match to known proteins was returned from sequence database searches (D. A. Bonocora & R. P. Shub, personal communication). With the exponential growth in the number of available sequences from the genome sequencing projects, increasing numbers of sequences cannot be aligned with certainty to known proteins on the basis of their sequence alone, and this limits the ability to assign a function to these sequences.

Functional identification using local sequence motifs

To overcome some of the problems associated with employing sequence alignments to determine protein function, several groups have developed databases of short sequence patterns or motifs designed to identify a given function or activity of a protein. These databases, notably Prosite (http://expasy.hcuge.ch/sprot/prosite.html; Bairoch *et al*., 1995), Blocks (http://www.blocks.fhcrc.org; (Henikoff & Henikoff, 1991) and Prints (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html; Attwood & Beck, 1994; Attwood & Beck, 1994; Attwood *et al*., 1994, 1997), use short stretches of sequence information to identify sequence patterns that are specific for a given function; thus, they avoid the problems arising from the necessity of matching entire sequences. Protein

function can be identified by either a single, local sequence motif or a set of local motifs. Typically, a local sequence pattern is developed by first identifying the functionally important residues from a literature search. A set of proteins that are known to belong to the family are aligned and, on this basis, the minimal local sequence signature is developed. This signature is then tested against the sequence database and, if false positives are found, the sequence alignment is used to identify conserved residues that are then added to the signature. This process is iterated until a local signature of some specificity is derived. The Prints and Blocks databases use multiple alignment representations to improve the specificity. Developers of the Blocks database have automated the procedure for producing patterns (Henikoff & Henikoff, 1991). Any newly determined sequence can be rapidly compared to these dictionaries of patterns in Prosite, Prints and Blocks, and if any matches are found, the new sequence can be assigned to the corresponding functional family. In practice, these approaches are quite successful. As a result of their utility and power, the Prosite, Prints and Blocks databases are regularly used by the scientific community.

Conservation of three-dimensional structure in protein active sites

While use of sequence signatures for protein function prediction is very powerful, they still fail to identify protein function for a variety of reasons, all of which, in principle, stem from the fact that the chemistry required for the functionality of protein active sites arises from their three-dimensional structure. Thus, as sequences diverge, only those residues required for the chemistry of the protein activity will be absolutely conserved. The structure of these active-site residues in three-dimensional space should also be conserved. In general, local sequence motifs will be unable to recognize such conserved three-dimensional structure, especially if it involves residues that are non-local in sequence. Although the Prints (Attwood *et al*., 1994) and Blocks (Henikoff & Henikoff, 1991) databases have attempted to circumvent this problem by developing multiple local sequence signatures for a given functional family, the three-dimensional structure of the active site is still not represented by these one-dimensional sequences. But, it is the three-dimensional structure of active site residues that is explicitly conserved, as illustrated by the following examples.

The three-dimensional structure of urease was recently compared to those of adenosine deaminase and phosphotriesterase (Holm & Sander, 1997b). Previous one-dimensional sequence comparison had failed to detect any relationships between these proteins; however, comparison of their three-dimensional structures showed conservation of local structure around the active site, although the global folds are different. This same active-site geometry was then observed in an even larger family of enzymes, with an even greater diversity of overall tertiary structure, that are involved in nucleotide metabolism (Holm & Sander, 1997b). The geometry of the active site would not be recognized by local sequence signatures or by overall comparison of global tertiary structures, but only from an analysis of the structure of the functional residues around the active site. In another example, an analysis of the ribonucleotide reductases from archaebacteria, eubacteria and eukaryotes shows that critical cysteine residues in the catalytic domain of this enzyme are conserved across all organismal boundaries (Tauer & Benner, 1997). However, once again based on sequence alignment alone, the ribonucleotide reductases are not obviously related.

The more divergent the sequences are, the more difficult it is to show a familial functional relationship just by sequence comparison, even if the catalytically important residues are invariant. At the limit, proteins with completely different structures can have similar functions. The bacterial and eukaryotic serine proteases, having very different protein structures and very similar active sites (Branden & Tooze, 1991), illustrate this point. Local sequence signatures would be unable to recognize these proteins as belonging to the same functional family because there would be no sequence similarity other than the identity and relative orientation of the specific active-site residues, which are non-local in sequence.

Thus, based on the above data, one must identify the global fold of a protein and the specific geometric arrangement of the active-site residues. In other words, one needs to determine both the global fold and the local structure of those residues that are functionally important. Local sequence signatures, although very powerful, may not be able to recognize the active-site residues, because sequence information is inherently one-dimensional, while protein active sites are inherently three-dimensional. But, a method based on identifying the conserved structure found in protein active sites could easily recognize the active-site residues and could classify such proteins as belonging to a given functional family.

The sequence-to-structure-to-function paradigm and its application to function prediction

In what follows, we describe such a method for identification of protein function based on the sequence-to-structure-to-function paradigm. We make the reasonable assumption that three-dimensional information is important to the chemistry of protein function; therefore, the active-site structure of the residues responsible for that function will be conserved and we can identify it. In this spirit, we develop three-dimensional descriptors of specific protein functions, termed fuzzy functional forms or FFFs, based on the geometry, residue identity,

and conformation of protein active sites. These FFFs are based on known crystal structures of members of the functional family and on experimental data available from the literature. The idea is similar in concept to that of Hellinga and Richards, who developed three-dimensional descriptors of metal-binding sites in order to introduce novel binding sites into proteins (Hellinga *et al.*, 1991; Hellinga & Richards, 1991). Instead of making the descriptors overly specific, however, we explore how much they can be relaxed (i.e. made ''fuzzy'') while still specifically identifying the correct active sites in a database of known structures. We then show that these fuzzy functional descriptors developed on the basis of protein models can identify protein active sites not only from experimentally determined structures, but also from predicted protein structures provided either by *ab initio* folding algorithms or by threading algorithms. Thus, low-to-moderate resolution structures produced by current structure prediction algorithms are sufficient to identify active sites in these models. These results should allow us to significantly extend the analysis of functional families further into and beyond the twilight zone of sequence similarity, and should allow a more extensive functional analysis of the rapidly expanding genomic databases.

Here, the disulfide oxidoreductase activity of the glutaredoxin/thioredoxin family and the RNA hydrolytic activity of the T_1 ribonuclease family are presented as illustrations and proof-of-principle of the method. First, however, to illustrate the need for a new approach, we discuss the problems arising when local sequence signatures are used to identify the disulfide oxidoreductase activity of the glutaredoxin/thioredoxin family. Next, we describe the development of the FFF for this activity and demonstrate its specificity in identifying active sites in exact protein models. We then show that the FFF can specifically identify active sites in low-to-moderate resolution models produced by either *ab initio* folding or threading algorithms. Based on the application of the glutaredoxin/thioredoxin FFF to a threading model, a prediction of a novel active site in the gamma subunit of yeast glycosyl transferase and prediction of the active sites for two hypothetical yeast proteins whose functions have not been previously identified by either the Prosite, Prints or Blocks databases are described. Finally, to demonstrate that the result is not exclusive to the glutaredoxin/thioredoxin family, we present some results for the RNA hydrolytic active site of the T_1 ribonuclease family.

Results

Analysis of the performance of local sequence motifs for identifying function

As mentioned in the Introduction, local sequence signatures designed for function identification become increasingly less specific as the number of sequences within a protein family increases. To illustrate this point more fully, we performed an analysis of the Prosite database (Release 13.0, November, 1995). All instances of true positive, false positive and false negative sequences, as identified by the Prosite developers, for each family were collected and the results are plotted in Figure 1. These data clearly demonstrate that local sequence signatures perform quite well on many families, especially when the number of sequences found in a family is low. Of the 1152 patterns in this release of Prosite, 908 (79%) of the patterns were specific for their sequences (using the set of true and false positives and negatives as identified by the Prosite developers). However, as the number of observed instances of a local pattern increases, the number of false positives also tends to increase. For 10.5% of the patterns, 90 to 99% of the selected sequences were true positives, while for the remaining 10.5% of the patterns, less than

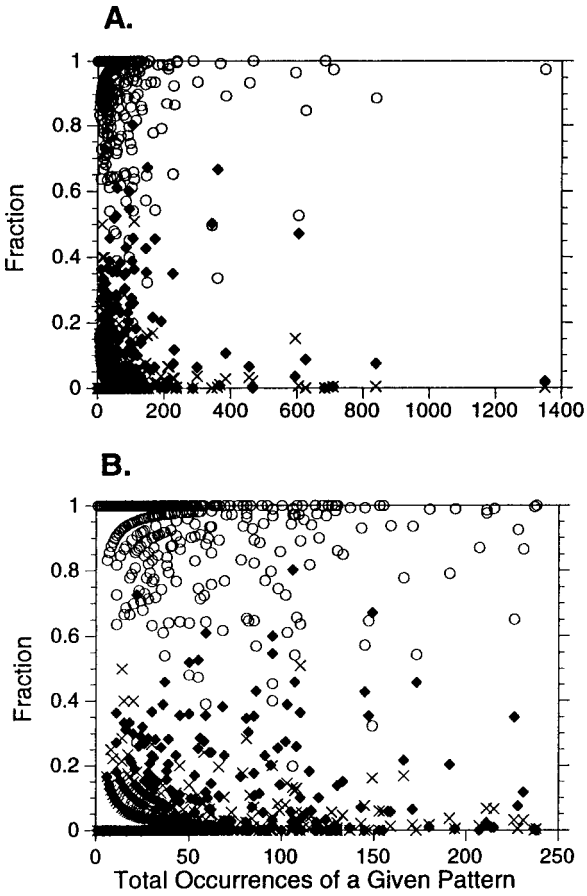


Figure 1. Data for 1152 patterns found in the Prosite database (Release 13.0). Fraction is the fraction of true positives (open circles), false positives (filled diamonds), or false negatives (X) found out of the total number of pattern occurrences. True positives, false positives and false negatives are those identified by the Prosite developers. A, All the data for 1152 patterns; B, the same data, with an expanded view of the *x*-axis.

90% of the selected sequences were true positives (Figure 1).

However, the results illustrated in Figure 1 are not the entire story. The data in Figure 1 were compiled only from those true and false positives and negatives identified by the Prosite developers. But identification of true and false positives and negatives can be ambiguous and can differ among the Prosite, Blocks and Prints databases. The current Prosite database (updated September 10, 1997) lists 111 true positives, five false positives and one false negative for its thioredoxin sequence signature (PS00194). The five false positives (YNC4_CAEEL, and POLG proteins from four poxyviruses) are not found by the thioredoxin sequence signature in either Blocks or Prints (Table 1). Three other proteins (FIXW_RHILE, GSBP_CHICK and RESA_BACSU) are identified as true positives in the Prosite database. Whether or not they are, in fact, truly thioredoxins, they are variously classified by Prints and Blocks (Table 1).

Database searches can reveal other sequences likely to belong to the thioredoxin family that are not listed in the sequence motif databases. For example, a keyword search of SwissProt (Bairoch & Apweiler, 1996) *via* the Sequence Retrieval System (SRS) at EMBL (http://www.embl-heidelberg.de/srs5) using the word ''thioredoxin'' revealed seven additional sequences (Table 1) that were identified as being thioredoxins or probable thioredoxins by the depositors of these sequences. These sequences are variously classified by Prosite, Prints and Blocks (Table 1). One sequence, Y039_MYCTU, is not recognized by any of these motif databases, demonstrating that some proteins possibly belonging to the thioredoxin family are not found by any of the local sequence signature databases. These results point out the need to enhance or improve the identification of function from protein sequence so as to be able to completely analyze the exponentially increasing genomic databases.

Finally, experimental evidence can suggest other proteins that might belong to the thioredoxin family. YME3_THIFE is a hypothetical 9.0 kDa protein in the MOBE 3′ region (ORF 8) in *Thiobacillus ferrooxidans*. A clone containing this gene is able to complement an *E. coli* thioredoxin mutant (Rohrer & Rawlings, 1992), thereby providing some experimental evidence that this hypothetical protein might fall into the glutaredoxin/thioredoxin family. A blast search of a non-redundant

Table 1. Classification of possible thioredoxin sequences by the Prosite, Prints and Blocks motif databases

	Sequence recognized by		
	Prosite	Prints	Blocks
A. *Sequences inconsistently classified by the three motif databases*			
FIXW_RHILE	X		X
GSBP_CHICK	X	X	X
RESA_BACSU	X	X(2)[a]	X
B. *Sequences found by keyboard search of SwissProt for ''thioredoxin''*			
DSBC_HAEIN			X
THIO_CHLLT		X(2)[a]	X
THIO_CHRVI		X	X
THIO_RHORU		X	
YX09_MYCTU		X	
Y039_MYCTU			
YB59_HAEIN			X
C. *Sequences with some experimental evidence*			
YME3_THIFE[b]			X
D. *False positives found by Prosite*			
YNC4_CAEEL	X		
POLG_PVYC	X		
POLG_PVYN	X		
POLG_PVYHU	X		
POLG_PVYO	X		

Prosite, recent Prosite database online; thioredoxin examples updated 9/10/97; http://expasy.hcuge.ch/cprot/prosite.html; Bairoch *et al.* (1995).

Prints, search of PWL26.0 database; http://www.biochem.ucl.ac.uk.bsm/dbbrowser/PRINTS/PRINTS.html; Bleasby *et al.* (1994).

Blocks, search of SwissProt32; http://www.blocks.fhcrc.org; Bairoch & Apweiler (1996).

[a] Prints uses three different sequence signatures to recognize the thioredoxins; (2) means that this sequence was recognized by only two of the three signatures.

[b] A plasmid in *E. coli* expressing this gene product complements a thioredoxin mutant, providing experimental evidence that this protein may be a glutaredoxin or thioredoxin (Rohrer & Rawlings, 1992).

sequence database (Genbank CDS translations, PDB, SwissProt and PIR; http://www.ncbi.nlm.nih.gov/BLAST/blast-databases.html) using YME3_THIFE as the search sequence produces two significant matches and two sequence twilight zone matches. The significant matches are to a periplasmic hydrogenase from *D. vulgaris* (PHFL_DESVO) and an open reading frame (ORF-R5) from *Anabaena*. One of the two twilight zone matches is to GLRX_METTH, a glutaredoxin-like protein from *Methanobacterium thermoautotrophicum*. GLRX_METTH itself exhibits significant sequence similarity to a number of thioredoxins. Examination of the sequence alignment between GLRX_METTH and YME3_THIFE shows that the active-site cysteine residues are conserved. Thus, this hypothetical protein, YME3_THIFE, is very weakly similar to GLRX_METTH and probably belongs to the glutaredoxin/thioredoxin family. Even though YME3_THIFE can be identified by weak sequence similarity and there is experimental evidence that it belongs to this family, the sequence is not identified as such by Prosite, because it contains only a portion of either the glutaredoxin or thioredoxin Prosite signatures (Figure 2). Furthermore, it is not found by Prints, but is classified as a glutaredoxin (because of the weak sequence similarity found by the BLAST alignment to GLRX_METTH) by Blocks (Table 1). These examples further illustrate the need to expand the ability to identify protein function from sequence, the power and utility of these sequence signature databases notwithstanding.

Development of a FFF for the disulfide oxidoreductase activity of the glutaredoxin/thioredoxin protein family

The glutaredoxin/thioredoxin protein family is composed of small proteins that catalyze thiol–disulfide exchange reactions *via* a redox-active pair of cysteine residues in the active site (Yang & Wells, 1991a,b). While glutaredoxins and thioredoxins catalyze similar reactions, they are distinguished by their differential reactivity. Glutaredoxins contain a glutathione-binding site, are reduced by glutathione (which is itself reduced by glutathione reductase), and are essential for the glutathione-dependent synthesis of deoxyribonucleotides by ribonucleotide reductase (Holmgren & Aslund, 1995). In contrast, thioredoxins are reduced directly by the specific flavoprotein thioredoxin reductase and act as more general disulfide reductases (Holmgren & Bjornstedt, 1995). Ultimately, however, reducing equivalents for both proteins come from NADPH. Protein disulfide isomerases (PDIs) have been found to contain athioredoxin-like domain and thus have a similar activity (Kemmink *et al*., 1995, 1997).

The active site of the redoxin family contains three invariant residues: two cysteines and a *cis*-proline. Mutagenesis experiments have shown that the two cysteine residues separated by two residues are essential for significant protein function. The side-chains of these two residues are oxidized and reduced during the reaction (Bushweller *et al*., 1992; Yang & Wells, 1991a). These two cysteine residues are located at the N terminus of an α-helix. Peptide studies have suggested that the positive end of the helix macrodipole affects the ionization of the cysteine residues and is thus conjectured to be important for protein function (Kortemme & Crieghton, 1995, 1996), although alternative views have been expressed (Dyson *et al*., 1997). Another unique feature of the redoxin family is the presence of a *cis*-proline residue located close to the two cysteine residues in structure, but not in sequence. While this proline residue is structurally conserved in all glutaredoxin and thioredoxin structures (Katti *et al*., 1995) and is invariant in aligned sequences of known glutaredoxins and thioredoxins, its functional importance is unknown. Other residues, particularly charged residues, have been shown to be important for the specific thiol characteristics of the cysteine residues, but are not essential and can vary within the family (Dyson *et al*., 1997).

The FFF for the disulfide oxidoreductase activity of the glutaredoxin/thioredoxin family was built as described in Methods and outlined in Figure 3. The literature information incorporated into the FFF is described above. The structure of the active site was taken from the three-dimensional structural comparison of bacteriophage T4 glutaredoxin, 1aaz (Eklund *et al*., 1992), human thioredoxin, 4trx (Forman-Kay *et al*., 1990) and disulfide bond formation protein, 1dsb (Martin *et al*., 1993). The superposition of the active sites of these three proteins is shown in Figure 4A, with the α-carbon distances between the relevant residues used to create the FFF shown in Table 2. The set of α-carbon dis-

```
YME3_THIFE:   K V  E      V    FSAG C  P     S   C  QT    A    IE      L
GLRX_METTH:       V      F    TSPT C  P     Y   C  PM    A    IE      V
Glrx:           [LIVD]-[FYSA]-x(4)-C-[PV]-[FYW]-C-x(2)-[TAV]-x(2,3)-[LIV]
Thx: [LIVMF]-[LIVMSTA]-x-[LIVMFYC]-[FYWSTHE]-x(2)-[FYWGTN]-C-[GATPLVE]-[PHYWSTA]-C-
     x(6)-[LIVMFYWT]
```

Figure 2. A comparison of the local sequences of GLRX_METTH, a ''glutaredoxin-like'' protein assigned to be a glutaredoxin by the Prosite sequence signature PS00195; YME3_THIFE, a protein not recognized by the Prosite or Prints sequence signature; and the Prosite sequence signature for the glutaredoxins (Glrx, PS00195). The sequences are aligned with the Glrx signature for easy comparison. The Prosite sequence signature for the thioredoxins (PS00194) is presented. It can be seen that the YME3-THIFE sequence does not match the thioredoxin signature either.

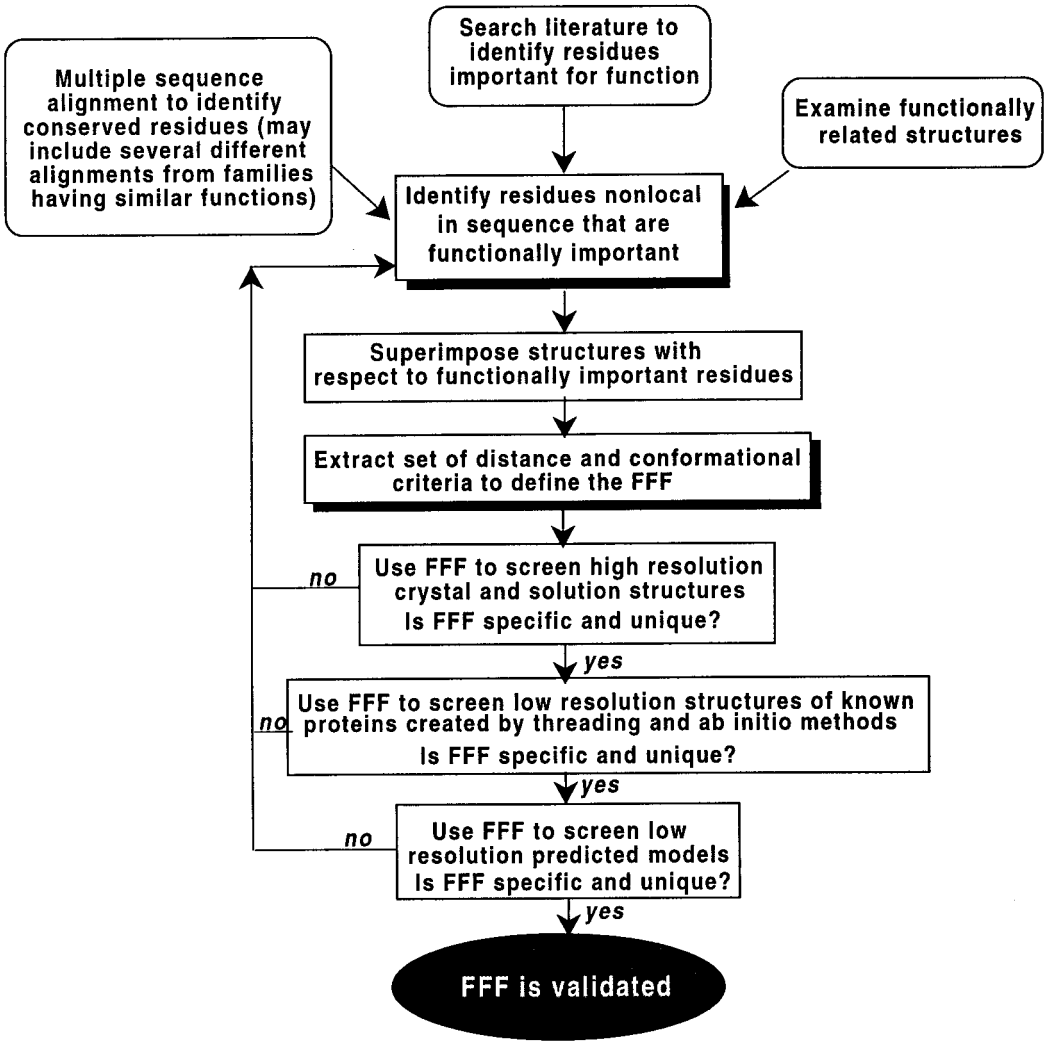
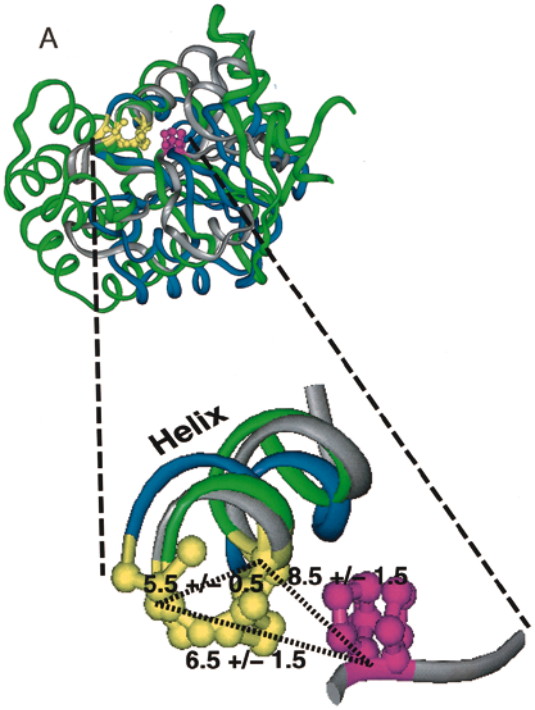


Figure 3. Outline of the protocol for producing a FFF. Information from library searches, multiple sequence alignments, and examination of tertiary structures are used to identify residues that are functionally important. From there, the FFF is defined and validated first on high-resolution crystal and solution structures, then on low-to-moderate resolution models of known structures, and finally on predicted models. At each of these steps, we ask whether the FFF is unique and specific for the given activity. If it is not, the active-site residues are re-evaluated and additional constraints are added to the FFF to make it more specific. This procedure was used to create the disulfide oxidoreductase and T_1 ribonuclease FFFs, as shown by the data presented in Table 2.

tances that define the FFF and their location with respect to the active site are indicated by dotted lines. The following FFF was thus developed: two cysteine residues separated by two residues and an α-carbon distance of 5.5 (±0.5) Å. These cysteine residues must be close to a proline residue. The α-carbon distance from Cys(i) to the proline residue is 8.5 (±1.5) Å and that from Cys($i + 3$) is 6.5 (±1.5) Å. These three sets of distances comprise the distances-only FFF of the glutaredoxin/thioredoxin family. The definition of the FFF itself and the specific geometric information used to create the FFF are shown in Table 2. There is some evidence that the cysteine residues must be at the N terminus of a helix because of the effect of the helix macrodipole on the sulfhydryl ionization (Kortemme & Crieghton, 1995, 1996); however, this evidence is disputed (Dyson *et al.*, 1997), so this characteristic is applied only if necessary.

Application of the glutaredoxin/thioredoxin FFF to exact protein structures

The distances-only FFF (Figure 4 and Table 2) is almost sufficient to uniquely distinguish proteins belonging to the glutaredoxin/thioredoxin family from a data set of 364 non-redundant proteins taken from the Brookhaven database. For this set of 364 proteins, 13 have the sequence signature-C-X-X-C-. Of these, only three, 1thx (thioredoxin), 1dsbA (protein disulfide isomerase, chain A) and 1prcM (photosynthetic reaction center, chain M) have a proline residue within the distances specified in Table 2. Of these three, only 1thx and 1dsb



B

```
           1                                                    50
GLR1_ECOLI ..........  ..........  ..........  ........MQ TVIPGRSGCP
 THIO_BPT4 ..........  ..........  ..........  ........MF KVYGYDSNIH
THIO_HUMAN ..........  ........VK QIESKTAFQE ALDAAGDKLV VVDFSATWCG
DSBA_ECOLI MKKIWLALAG LVLAFSASAA QYEDGKQYTT LEKPVAGAPQ VLEFFSFFCP

           51                                                   100
GLR1_ECOLI YCVRAKDLAE KLSNERDDF. ...QYQYVDI RAEGITKE.. .DLQQKAGKP
 THIO_BPT4 KCVYCDNAKR LLTVKKQPF. ...EFINIMP EKGVFDDEKI AELLTKLGRD
THIO_HUMAN PCKMIKPFFH SLSE...KYS N.VIFLEVDV D......... ..DCQDVASE
DSBA_ECOLI HCYQFEEVLH ISDNVKKKLP EGVKMTKYHV NFMGGDLGKD LTQAWAVAMA

           101                                                  150
GLR1_ECOLI VE...TVPQI FV.DQQHIGG YTDFAAWVKE NLDA...... ..........
 THIO_BPT4 TQIGLTMPQV FAPDGSHIGG FDQLREYFK. .......... ..........
THIO_HUMAN CEVKCMPTFQ FFKKGQKVGE .......... .FSGA.NKEK LEATINELV.
DSBA_ECOLI LGVEDKVTVP LFEGVQKTQT IRSASDIRDV FINAGIKGEE YDAAWNSFVV

           151                                                  200
GLR1_ECOLI ..........  ..........  ..........  ..........  ..........
 THIO_BPT4 ..........  ..........  ..........  ..........  ..........
THIO_HUMAN ..........  ..........  ..........  ..........  ..........
DSBA_ECOLI KSLVAQQEKA AADVQLRGVP AMFVNGKYQL NPQGMDTSNM DVFVQQYADT

           201
GLR1_ECOLI ........
 THIO_BPT4 ........
THIO_HUMAN ........
DSBA_ECOLI VKYLSEKK
```

Figure 4. A, Structure of the proteins used to describe the disulfide oxidoreductase FFF, T4 glutaredoxin, laaz, chain A (Eklund *et al*., 1992; gray ribbon), human thioredoxin, 4trx (Forman-Kay *et al*., 1990; blue ribbon) and proline disulfide isomerase, 1dsb, chain A (Martin *et al*., 1993; green ribbon), and an enlargement of the active site of these proteins. The enlargement shows that the active-site structure of these proteins is conserved, although the structure of the rest of the proteins is very different. The backbone atoms of the two cysteine residues and the *cis*-proline residue were superimposed and the side-chains of the two cysteine residues (yellow) and the proline residue (magenta) are shown as black ball and stick models. The helix containing the cysteine residues is shown in orange and the N-terminal turn of this helix is cyan. The two residues on either side of the proline residues are shown as a gray ribbon. The C^{α}–C^{α} distances taken from these proteins used to define the FFF are shown in Table 2 and are indicated in the Figure by dotted lines. B, Sequence alignment of the laaz (THIO_BPT4), 1dsb, chain A (DSBA_ECOLI) and 4trx (THIO_HUMAN), the three proteins used to create the FFF, and 1ego (GLRl_ECOLI), the protein whose structure was predicted using the MONSSTER *ab initio* folding algorithm, The two cysteine residues and the *cis*-proline residue involved in the active site and specifically selected by the FFF are shown in bold and underlined. The Figure was created using Pileup (Winsconsin Package, Version 8, 1994, Genetics Computer Group). Comparison of A and B shows that the active-site residues in these proteins are conserved in the structure, but are not close in the sequence.

have their cysteine residues positioned at or near the N terminus of a helix. These two proteins are the only two ''true positives'' in the test data set, thereby showing that this simple FFF is quite specific for identifying the disulfide oxidoreductase active site of the glutaredoxin/thioredoxin protein family. When the requirement that the cysteine residues be at the N terminus of a helix is included, then the 1prc-M site is also eliminated, making the FFF absolutely specific for the glutaredoxin/thioredoxin disulfide oxidoreductase FFF.

To explore the ''fuzziness'' of this active-site descriptor, the allowed variance in the Cys-Pro and Cys-Cys α-carbon distances was uniformly increased in increments of ±0.1 Å. Upon increasing the allowed distances by ±0.1 Å, 1fjm (Goldberg *et al*., 1995), a serine/threonine phosphatase, 1lct (Day *et al*., 1993), a lactoferrin, and 1prc-C (Deisenhofer *et al*., 1995), the C-chain of the photosynthetic reaction center, were also selected by the distances-only FFF. The Cys-Cys-Pro site in 1fjm is curiously similar to that found in the glutaredoxin/thioredoxin family, including the proline residue being in a *cis*-conformation, but the cysteine residues are at the C terminus, not the N terminus, of a helix. 1lct, an iron transport protein, contains a proline residue near a cluster of metal-binding cysteine residues and these are in a very irregular structure, not in a helix. In 1prc-M, the Cys-Cys-Pro structural motif is located along one face of a transmembrane helix, near the C terminus of that helix. In 1pcr-C, the Cys-Cys-Pro motif is located in another very irregular region. Thus, even when relaxed or fuzzy descriptors are used, all four proteins found by the distances-only FFF are eliminated when the helix requirement is included. When the distance constraints are relaxed even further to ±0.3 Å, only one other protein, 2fd2 (Soman *et al*., 1991), a ferredoxin, is selected. Ferredoxin is another metal-binding protein. Again, the cysteine residues are found in a non-regular structural region, not in a helix. It is important to note that when the cysteine residues are required to be at the N terminus of a helix, all of these false positives are no longer recognized, even when the FFF distance constraints are further relaxed by ±0.3 Å from the distances and their allowed variances shown in Table 2.

Table 2. The distance data used to derive each FFF, a description of each FFF, and comparison to test structures

Residues involved in FFF		Distance in training proteins					Mean	SD	FFF definition *FFF-DIST-*	*FFF-Var-*	Test structure
A. *Disulfide oxidoreductase FFF of the glutaredoxins/thioredoxins*											
		1aazA	1aazB	1dsbA	1dsbB	4trx					1ego
Cys(*i*)	Pro	7.93	8.15	7.47	7.70	10.55	8.360	1.25028	8.5	1.5	7.48
Cys(*i*+3)	Pro	5.01	5.17	5.61	5.28	6.27	5.468	0.49932	6.5	1.5	5.60
Cys(*i*)	Cys(*i*+3)	5.27	5.21	5.18	5.55	5.54	5.350	0.18097	5.5	0.5	5.33
B. *RNA hydrolytic FFF of the T_1 ribonucleases*											
		1rtu	1fus	1rms							9rnt
His	His*	15.20	16.70	15.97			15.950	0.6577	15.9	1.5	15.63
His	Glu	5.36	5.84	5.71			5.637	0.2221	5.7	0.5	5.79
His*	Glu	13.03	12.90	21.44			12.580	0.2773	12.6	1.0	11.95
Tyr	Phe	13.03	16.40	16.62			16.580	0.1290	16.5	0.5	16.43
Tyr	Arg	10.50	10.20	10.25			10.330	0.1473	10.3	0.5	10.29
Phe	Arg	9.61	9.34	9.40			9.450	0.1418	9.5	0.5	9.59
His	Tyr	4.87	5.02	5.13			5.007	0.1305	5.0	0.5	5.07
His	Phe	14.47	15.60	15.28			15.120	0.5866	15.2	1.0	15.28
His	Arg	10.44	11.30	10.94			10.900	0.4366	11.0	1.0	11.16
His*	Tyr	16.06	16.10	15.86			16.010	0.1286	15.80	1.0	15.32
His*	Phe	4.67	4.60	4.63			4.633	0.0351	4.6	0.5	4.64
His*	Arg	8.72	8.79	8.50			8.670	0.1513	8.6	0.5	8.48
Glu	Tyr	7.36	7.10	7.13			7.197	0.1422	7.2	0.5	7.24
Glu	Phe	21.17	11.80	11.77			11.900	0.2309	11.9	0.5	11.96
Glu	Arg	6.33	6.16	5.87			6.120	0.2326	6.1	0.5	6.00

The residues in each family used to define the FFF are shown in Figures 4A and 7B. In the glutaredoxin/thioredoxin family, two cysteine residues separated by two residues and a proline residue are used. In the T_1 ribonucleases, six residues are used: the nucleophilic triad consisting of two histidine residues, His and His*, with the former (latter) closer to the N(C) terminus and a glutamic acid residue, and the transition state for stabilization triad consisting of a tyrosine, an arginine and a large hydrophobic residue (phenylalanine in 1rtu, 1fus and 1rms). For the T_1 ribonucleases FFF, the first three lines are the nucleophilic triad; the next three lines are for the transition state stabilization triad; and the last nine lines are the distances between the two triads. Exact C^{α}–C^{α} distances for the relevant residues in the proteins used to define each FFF are presented (1aazA, 1aazB, 1dsbA, 1dsbB and 4trx for the glutaredoxin/thioredoxin family, and 1rtu, 1fus and 1rms for the T_1 ribonuclease family). Mean is the mean of the α-carbon distances (in Å) found in these structures; SD is the standard deviation for this distribution of distances. The columns *FFF-DIST* and *FFF-Var* are the data that describe the FFF: *DIST* is the C^{α}–C^{α} distance and *Var* is the variation allowed in these interatomic distances. In most cases, *DIST* is close to the mean distance found in proteins used to define the FFF (and is derived using the set of 364 PDB structures and is chosen to eliminate false positives, and *Var* is correlated with the standard deviation for the distribution of distances found in the same set. In the last column, the distances for (A) 1ego (Xia *et al*., 1992) and (B) 9rnt (Martinez-Oyanedel *et al*., 1991), the test proteins that were not used to define the FFF but were used to test its definition, are given for comparison.

Application of the glutaredoxin/thioredoxin FFF to predicted protein models produced by an *ab initio* folding algorithm

Current *ab initio* protein structure prediction algorithms can often generate inexact models of proteins or protein fragments with a 3 to 6 Å backbone coordinate root-mean-square deviation (cRMSD) from the native structure (Aszodi *et al*., 1995; Friesner & Gunn, 1996; Mumenthaler & Braun, 1995; Ortiz *et al*., 1998; Smith-Brown *et al*., 1993; Srinivasan & Rose, 1995). Is the glutaredoxin/thioredoxin FFF sufficient to identify the active site of such an inexact model of a protein, or is a high-resolution crystal or solution structure required? The structure of *E. coli* glutaredoxin, 1ego (Xia *et al*., 1992), was predicted with a 5.7 Å cRMSD of the α-carbon atoms by the MONSSTER algorithm (Skolnick *et al*., 1997). Furthermore, the sequence of this glutaredoxin exhibits less than 30% sequence identity with any of the three structures used to create the FFF (Figure 4B). The disulfide oxidoreductase FFF was applied to 25 ''correct'' structures and 56 ''incorrect'' or ''misfolded'' structures generated by MONSSTER for the 1ego sequence during the isothermal runs. The distances-only FFF specifically selected all 25 ego-like structures as belonging to the redoxin family and rejected all 56 misfolded structures. A set of 267 correctly and incorrectly predicted structures produced by the MONSSTER algorithm for five other proteins was then created. The distances-only glutaredoxin/thioredoxin FFF was specific for the correctly folded ego structures and did not recognize any of the correctly or incorrectly folded structures of these other proteins. Inclusion of the more specific criterion that the cysteine residues be at the N terminus of a helix did not change these results.

To further explore the allowed fuzziness of the FFF as applied to these inexact models, the distance constraints were again relaxed. When the allowed variance in the α-carbon distances shown in Table 2 was relaxed by an additional ±0.2 Å, the FFF was still absolutely specific for all correctly folded 1ego structures. When the variance was relaxed to ±0.3 Å, the distances-only FFF picked up two of the 56 misfolded 1ego structures, in

addition to the 25 correctly folded structures. When the allowed variance was further relaxed to ±0.5 Å, no additional incorrectly folded structure was selected. These results demonstrate the specificity and the uniqueness of the glutaredoxin/thioredoxin disulfide oxidoreductase FFF for low resolution models of protein structure by *ab initio* folding algorithms.

Application of the glutaredoxin/thioredoxin disulfide oxidoreductase to models built using threading or inverse folding algorithms

The FFF concept can be applied to proteins that have been folded by an *ab initio* folding algorithm, but such state-of-the-art algorithms are too slow to permit genome-wide screening. Thus, for large-scale screening, it would be most useful if the FFFs could be applied to three-dimensional protein models produced by threading or inverse folding algorithms. To be useful for genome-wide screening, the procedure must recognized proteins that could not be detected by standard sequence analysis methods. Thus, we applied the disulfide oxidoreductase FFF to several putative proteins from the yeast genome database (Mewes *et al*., 1997). The selected protein sequences were aligned to a database of 301 non-homologous protein structures (Fischer *et al*., 1996) using an inverse folding or threading algorithm (Jaroszewski *et al*., 1998). Models were built using automatic scripts, as described in Methods. Without further relaxation, these models were screened using the glutaredoxin/thioredoxin FFF.

A total of eight ORFs from the *S. cerevisiae* genome database were tested (Table 3). Six were identified by the combination of threading and FFF as containing the disulfide oxidoreductase active site: one protein is predicted to belong to the protein disulfide isomerase family (S67190); one sequence that the depositors identify as a hypothetical thioredoxin (YCX3_YEAST); one sequence, which has no detectable sequence similarity to any glutaredoxin or thioredoxin, identified as the gamma subunit of glycosyl transferase (OSTG_YEAST); and three hypothetical proteins, one having very distant sequence similarity to glutaredoxin from rice (S51382), one with very distant sequence similarity (insignificant by the Blast score) to the glutaredoxin from *Methanococcus thermoautoformicum* (S70116), and one with no similarity to any glutaredoxin or thioredoxin by Blast (YBR5_YEAST). Of these six, only YCX3_YEAST, S67190, and S70116 were identified as a glutaredoxin or thioredoxin by at

Table 3. Results of the application of threading and the glutaredoxin/thioredoxin disulfide oxidoreductase FFF to eight sequences from the yeast genome database

Sequence	Blast	PS	P(PS)	P(B)	B	Aligns	Signif.	Active-site res.	Name
YCX3_YEAST	X	X	X	X	X	2trxA	48.1	C55,C58,P98	Hypothetical thrx-like protein
						2trxA	498.3		
						2trxA	101.2		
S67190	X	X	X	X	X	2trxA	5464.9	C59,C62,P105	MPD1 prot
						2trxA	1736.8		
						2trxA	2200.6		
S70116	X				X	1ego	7.6	C31,C34,P79	Hypothetical protein
						2trxA	17.6		
						2trxA	13.5		
S51382	X					1ego	7.8	C25,C28,P74	Hypothetical protein
						2trxA	17.5		
						2trxA	16.0		
YBR5_YEAST						1dsbA	29.0	C13,C16,P151	Hypothetical protein
OSTG_YEAST						2trxA	25.6	C73,C76,P133	Glycosyl transferase γ subunit
						2trxA	43.7		
						2trxA	95.8		
YEO4_YEAST	X					2trxA	175.6	NF	Hypothetical protein
						2trxA	934.3	NF	
						2trxA	730.3	NF	
YPRO82c	X					2trxA	8.9	NF	Hypothetical protein
						2trxA	23.1	NF	
						2trxA	23.3	NF	

The first five columns show if the function of the sequence could have been determined by sequence alignment or by the motifs databases: Blast (significant score by the gapped or Psi-Blast algorithm); PS, Prosite; P(PS), Prints scored by the Prosite method; P(B), Prints scored by the Blocks method; B, Blocks. Aligns is the structure to which the threading algorithm aligned the sequence. Signif. is the significance score of the alignment. The first entry of each group of three is the sequence-based scoring method; the second entry is the sequence-structure scoring method; the third is the structure-structure scoring method (Jaroszewski *et al*., 1998; and see Methods for a brief description). Note that YBR5_YEAST was aligned with 1ego, 2trxA or 1dsbA by only one scoring method; thus, only one score is given. Active-site res. are the residues in the threading model that are identified by the disulfide oxidoreductase glutaredoxin/thioredoxin FFF as being active site residues. NF means that, although the sequence aligned with 1ego, 2trxA or 1dsbA, the FFF did not identify a disulfide oxidoreductase active site in this protein.

least one of the motif databases. Two additional sequences, YEO4_YEAST and YPRO82c, were identified by the threading algorithm as having the glutaredoxin or thioredoxin structure, but were not identified by the FFF as containing the oxidoreductase active site. Both of these sequences exhibit identifiable sequence similarity to glutaredoxins or thioredoxins by the Blast algorithm (Table 3).

The threading algorithm (Jaroszewski *et al.*, 1998) aligns the sequences of all eight of these ORFs to the structure of either 1ego (*E. coli* glutaredoxin (Xia *et al.*, 1992)), 2trx, chain A (*E. coli* thioredoxin (Katti *et al.*, 1990)), or 1dsb, chain A (*E. coli* protein disulfide isomerase (Martin *et al.*, 1993)) from a database of 301 non-homologous proteins (Fischer *et al.*, 1996). The alignment fit is strong, as seven of the eight sequences were matched to 1ego, 2trx or 1dsb by all three scoring methods used to assess the significance of the threading results (Table 3). For comparison, the significance scores reported in Table 3 can be compared to the distribution of significance scores for all yeast sequences that aligned to 2trx, chain A (Figure 5). Models were built based on the sequence-to-structure alignments and were screened with the FFF. Sixteen of the models (one model for each scoring method for YCX3_YEAST, S67190, S70116, S51382, YBR5_YEAST and OSTG_YEAST) were found to have the disulfide oxidoreductase active site described by the distances-only FFF. The residues predicted to be in the active sites of these proteins are listed in Table 3.

This result is remarkable when one considers that the sequence similarity between these proteins is virtually non-existent. Several examples are shown in Figure 6. In one case (S70116), standard multiple sequence alignments even fail to correctly align the proposed active-site residues when the sequences are aligned with each other (Figure 6B). The Prosite, Prints and Blocks motif databases variously classify these proteins (Table 3). One sequence (S570116) is recognized only by the Blocks database; another sequence (S51382) is not recognized by any of these sequence motif databases. Furthermore, BLAST sequence alignment algorithms do not find a match for YBR5_YEAST or OSTG_YEAST to any glutaredoxins or thioredoxins. Thus, the result for these two sequences stands as a prediction of activity based on the disulfide oxidoreductase glutaredoxin/thioredoxin FFF.

Finally, it must be shown that not all sequences recognized by the threading algorithm contain the disulfide oxidoreductase active site. In theory, threading algorithms should recognize structure only; consequently, they should be able to recognize proteins with similar structures, but that do not have the same function. Such proteins are termed topological cousins. The FFF should allow us to distinguish between functionally related proteins and topological cousins. Two of the sequences (YEO4_YEAST and YPRO82c) are found to align with 2trx, chain A, by all three scoring methods used by the threading algorithm (Table 3). Com-

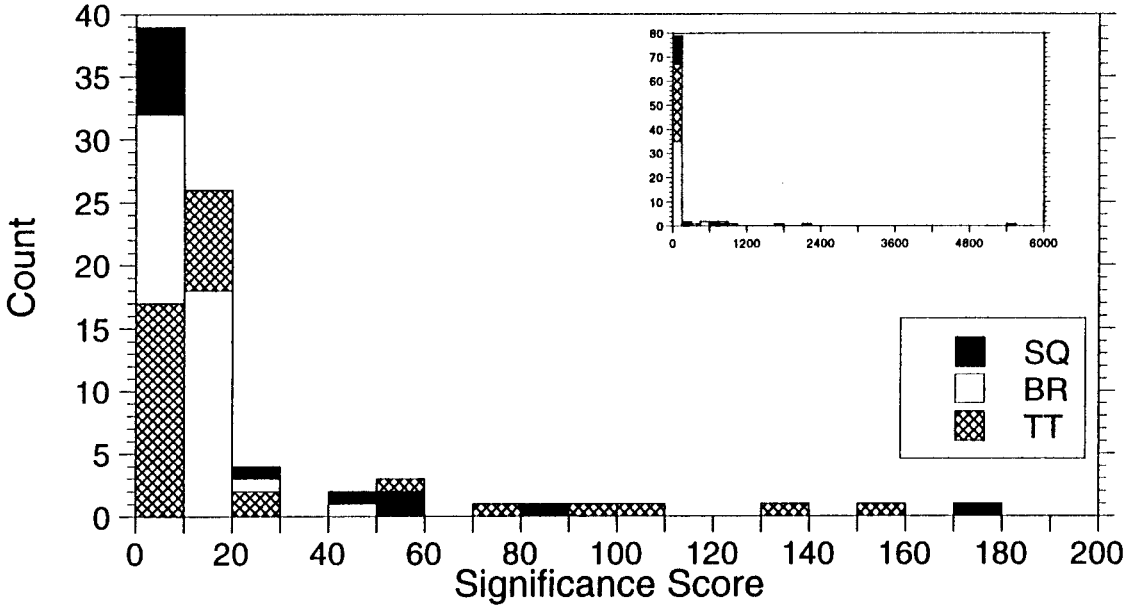


Figure 5. The distribution of significance scores for all sequences from the yeast genome that aligned with 2trx, chain A, by the threading algorithm. The main graph shows the distribution for significance scores of 1 to 200; the inset graph shows the distribution for all significance scores. Comparison of scores presented in Table 3 and this histogram shows that sequences found to contain the disulfide oxidoreductase active site and those that do not contain the active site have similar significance scores. Thus, application of the FFF allows us to automatically distinguish between proteins with similar active sites and those that are topological cousins. The threading algorithm and three different scoring methods (filled, open and cross-hatched bars) are described in detail by Jaroszewski *et al.* (1998) and briefly described in Methods.

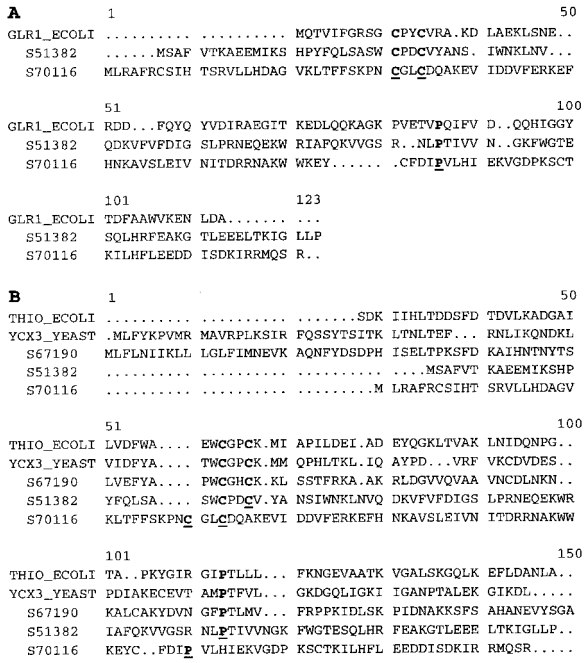


Figure 6. Sequence alignment of four proteins from the *Saccharomyces cerevisiae* genome with each other and with the target structures 1ego (Xia *et al*., 1992; GLRl_ECOLI) and 2trx (Katti *et al*., 1990; THIO_ECOLI; see Table 3. The threading program matched S51382 and S70116, both hypothetical proteins, to 1ego (A); the threading program matched YCX3_YEAST, a hypothetical thioredoxin-like protein, S67190, a protein that is predicted to be related to the protein disulfide isomerases, S51382 and S70116 to 2trx (B). (Different scoring functions matched S51382 and S70116 to different templates, thus they are shown in both alignments; see Table 3). In B, only the first 150 residues of S67190 that align with 2trx are shown. The cysteine and proline residues identified by the FFF as being part of the disulfide oxidoreductase active site (see Figure 4A and Table 2) are shown in bold type. Alignments were produced by the Pileup multiple sequence alignment program (Wisconsin package, Version 8, 1994, Genetics Computer Group) using the standard parameters.

parison of the scores reported in Table 3 to the distribution of scores shown in Figure 5 demonstrates that these are relatively strong structural predictions. Furthermore, both of these putative proteins are found to align to thioredoxins by the BLAST algorithm (Table 3), although the significance score of these sequence alignments is not high. However, the disulfide oxidoreductase FFF does not find the correct active site in the three-dimensional models. Analysis of the sequences by hand after the fact demonstrates that neither sequence has the CXXC sequence characteristic of the active site of this family, thus it is unlikely that these proteins would demonstrate the oxidoreductase activity. We have thus shown that not all proteins that align with 1ego, 2trx or 1dsb by the threading algorithm exhibit the disulfide oxidoreductase active site.

Taken together, these results demonstrate that models produced by threading algorithms are sufficient for application of the FFF to the identification of active sites in proteins. Application of the FFF idea is an automatic method for distinguishing between functionally related proteins and topological cousins. These results suggest a means for large-scale functional analysis of the genome databases using the sequence-to-structure-function paradigm.

Development of a FFF for the T_1 ribonuclease protein family

To show that the FFF concept is applicable to other active sites besides the glutaredoxin/thioredoxin disulfide oxidoreductase active site, a FFF was built for the active site of the T_1 ribonuclease family. The T_1 ribonucleases are a family of proteins that include a number of ribonucleases such as T_1, T_2, U_2 and F_1, and the distantly related family of fungal ribotoxins. These proteins are endoribonucleases that are generally specific for purine, particularly guanine, bases (Steyaert, 1997).

The catalytic mechanism of the T_1 ribonucleases has been well studied (for a review, see Steyaert, 1997). Two histidines and a glutamic acid residue are essential for the nucleophilic displacement of the phosphate atoms. A tyrosine, a phenylalanine (or another large hydrophobic residue) and an arginine residue are responsible for stabilizing the transition state of the reaction. These catalytic residues are located on various strands across one face of a β-sheet. They are highlighted in the multiple sequence alignment of T_1 ribonucleases shown in Figure 7A, and their proximity in three-dimensional space is shown in Figure 7B. Neither Prosite, Prints nor Blocks provides a sequence signature with which to identify this family.

An analysis of three T_1 ribonucleases whose structures have been solved (1rms, Nonaka *et al*., 1993; 1fus, Vassylyev *et al*., 1993; and 1rtu, Noguchi *et al*., 1995) shows that the location of the active-site residues in three-dimensional space is very well conserved. Thus, a FFF based on the distances between appropriate α-carbon atoms was developed from these distances, plus or minus a small variance. The exact values used to create the FFF and the distances and variances for the FFF itself are given in Table 2.

Application of the T_1 ribonuclease FFF to "exact" protein structures

When applied to three-dimensional structures, the T_1 ribonuclease FFF was implemented in three stages: first the structure is searched for the residue triad involved in nucleophilic displacement (His-His-Glu); second, the structure is searched for the residue triad involved in transition state stabilization (Tyr-Hydrophobic-Arg); third, if both triads

A

```
1                                                  50
RNT1_ASPOR ..........  ...MMYSKLL TLTTLLLPTA LALPSLVERA CDYTCGSNCY
RNF1_FUSMO ..........  .......... .......... .........Q SATTCGSTNY
RNMS_ASPSA ..........  .......... .......... ........ES CEYTCGSTCY
RNU2_USTSP ..........  .......... .......... ......CNIP ESTNCGGNVY
RNC2_ASPCL ..........  .......... .......... .........D CDYTCGSHCY
RNPB_PENBR ..........  .......... .......... .........A CAATCGTVCY
RNPC_PENCH ..........  .......... .......... .........A CAATCGSVCY
RNN1_NEUCR ..........  .......... .......... .........A CMYICGSVCY
RNU1_USTSP ..........  .......... .......... .......QGG VSVNCGGTYY
RNAS_ASPGI MVAIKNLVLV ALTAVTALAV PSPLEARAVT WTCLNDQKNP KTNKYETKRL
RNCL_ASPCL MVAIKNLVLV ALTAVTALAM PSPLEERAAT WTCMNEQKNP KTNKYENKRL
RNMG_ASPRE MVAIKNLFLL AATAVSVLAA PSPLDARA.T WTCINQQLNP KTNKWEDKRL

          51                                                  100
RNT1_ASPOR SSSDVSTAQA AGYQLHEDGE TVGSNSYPHK YNN.YEG... .......FDF
RNF1_FUSMO SASQVRAAAN AACQYYQNDD SAGSTTYPHT YNN.YEG... .......FDF
RNMS_ASPSA WSSDVSAAKA KGYSLYESGD TI..DDYPHG YHD.YEG... .......FDF
RNU2_USTSP SNDDINTAIQ GA...LDDVA RPDGDNYPHQ YYD.EAS... .......EDI
RNC2_ASPCL SASAVSDAQS AGYQLESAGQ SVGRSRYPHQ YRN.YEG... .......FNF
RNPB_PENBR TSSAISSAQA AGYNLYSTND DV..SNYPHE YHN.YEG... .......FDF
RNPC_PENCH TSSAISAAQE AGYDLYSAND DV..SNYPHE YRN.YEG... .......FDF
RNN1_NEUCR SSSAISAALN KGYSYYEDGA TAGSSSYPHR YNN.YEG... .......FDF
RNU1_USTSP SSTQVNRAIN NA.....KSG QYSSTGYPHT YNN.YEG... .......FDF
RNAS_ASPGI LYNQNKAESN SHHAPLSDGK T..GSSYPHW FTNGYDGDGK LPKGRTPIKF
RNCL_ASPCL LYNQNNAESN AHHAPLSDGK T..GSSYPHW FTNGYDGDGK ILKGRTPIKW
RNMG_ASPRE LYSQAKAESN SHHAPLSDGK T..GSSYPHW FTNGYDGNGK LIKGRTPIKF

          101                                                 150
RNT1_ASPOR ..S.VSSP.. .......... ..YYEWPILS SGDVYS..G. ...GSPGADR
RNF1_FUSMO ..P.VDGP.. .......... ..YQEFPIKS GG.VYT..G. ...GSPGADR
RNMS_ASPSA ..P.VSGT.. .......... ..YYEYPIMS DYDVYT..G. ...GSPGADR
RNU2_USTSP TLCCGPGS.. .......... ..WSEFPLVY NGPYYS..SR DNYVSPGPDR
RNC2_ASPCL ..P.VSGN.. .......... ..YYEWPILS SGSTYN..G. ...GGPGADR
RNPB_PENBR ..P.VSGT.. .......... ..YYEFPILK SGKVYT..G. ...SSPGADR
RNPC_PENCH ..P.VSGT.. .......... ..YYEFPILR SGAVYS..G. ...NSPGADR
RNN1_NEUCR ..P.TAKP.. .......... ..WYEFPILS SGRVYT..G. ...GSPGADR
RNU1_USTSP S.DYCDGP.. .......... ..YKEYPLKT SSSGYT..G. ...GSPGADR
RNAS_ASPGI GKSDCDRPPK HSKDGNGKTD HYLLEFPTFP DGHDYKFDSK KPKENPGPAR
RNCL_ASPCL GNSDCDRPPK HSKNGDGKND HYLLEFPTFP DGHQYNFDSK KPKEDPGPAR
RNMG_ASPRE GKADCDRPPK HSQNGMGKDD HYLLEFPTFP DGHDYKFDSK KPKEDPGPAR

          151                                 182
RNT1_ASPOR VVFNENNQ.L AGVITHTGAS G.NNFVECT. ..
RNF1_FUSMO VVINTNCE.Y AGAITHTGAS G.NNFVGCSG TN
RNMS_ASPSA VIFNGDDE.L AGVITHTGAS G.DDFVACSS S.
RNU2_USTSP VIYQTNTGEF CATVTHTGAA SYDGFTQCS. ..
RNC2_ASPCL VVFNDNDE.L AGLITHTGAS G.DGFVACY. ..
RNPB_PENBR VIFNDDDE.L AGVITHTGAS G.NNFVACT. ..
RNPC_PENCH VVFNGNDQ.L AGVITHTGAS G.NNFVACD. ..
RNN1_NEUCR VIFDSHGN.L DMLITHNGAS G.NNFVACN. ..
RNU1_USTSP VVYDSNDGTF CGAITHTGAS G.NNFVQCSY ..
RNAS_ASPGI VIYTYPNKVF CGIIAHTKEN Q.GELKLCSH ..
RNCL_ASPCL VIYTYPNKVF CGIVAHTREN Q.GDLKLCSH ..
RNMG_ASPRE VIYTYPNKVF CGIVAHQRGN Q.GDLRLCSH ..
```

B

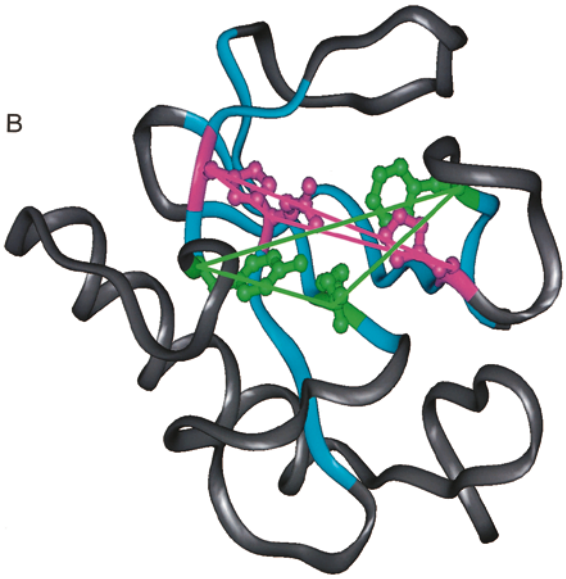


Figure 7. A, Sequence alignment of nine ribonucleases and three ribotoxins in the T_1 family. The first nine sequences are ribonucleases; the last three are ribotoxins with their leader sequences attached (RNAS_ASPGI, α-sarcin; RNCL_ASPCL, clavin; RNMG_ASPRE, restrictocin). The six residues involved in the catalytic mechanism are shown in bold. The first four sequences are those found in the PDB database and were used to create or to test the FFF (see Tables 2 and 4). The other T_1 ribonucleases were found by searching the SwissProt32 database (Bairoch & Apweiler, 1996) with BLAST (Altschul *et al.*, 1990) using RNT1_ASPOR as the search sequence. The ribonucleases were selected as the most significant matches to RNT1_ASPOR. All sequences, both ribonucleases and ribotoxins, were aligned using the Pileup multiple sequence alignment tool from the Wisconsin GCG package. B, A view looking along the approximate plane of the β-sheet that contains the active site of 9rnt (Martinez-Oyanedel *et al.*, 1991). The His-His-Glu involved in nucleophilic displacement are shown as magenta ball and stick models; the Tyr-Hydrophobic-Arg side-chains involved in transition state stabilization are shown as light green ball and stick models. The strands of the sheet (as identified by the crystallographer) are shown as a light blue ribbon. The set of α-carbon distances that define the FFF corresponding to the His-His-Glu distances are indicated by the magenta lines and the corresponding distances for those residues involved in transition state stabilization are shown by green lines. Comparison of A and B shows that the active-site residues are close in three-dimensional space, but not close in the one-dimensional sequence.

are found, the relative positions of the two triads are checked based only on the distances between α-carbon atoms. Application of the FFF to the 364 non-homologous PDB protein structures yields only one structure that contains both residue triads in the correct juxtapositions: 9rnt (Matinez-Oyanedel *et al.*, 1991), the only true positive in the test data set. Increasing the allowed variation for each distance by ±0.5 Å yields no additional protein, demonstrating that this FFF is specific for structures of the T_1 ribonuclease family solved to atomic resolution, even when the distance restraints are made increasingly fuzzy.

Application of the T_1 ribonuclease FFF on low-to-moderate resolution protein models

To test the applicability of the T_1 ribonuclease FFF to inexact, predicted models, the nine ribonuclease sequences listed in Table 4 and Figure 7A were threaded through 301 non-homologous proteins, as described in Methods. All nine sequences were matched as the highest score to the 9rnt structure by all three scoring methods. Models were built for all 27 (nine sequences times three scoring methods) sequence-to-structure alignments and all 27 models were screened by the T_1 ribonuclease FFF. All 27 models were found to contain both T_1 ribonuclease active-site triads in the correct locations in the structure (Table 3).

To test the method on more distantly related sequences, models of three ribotoxin sequences were built. Ribotoxins are a small family of proteins found in the *Aspergillus* fungi family. They cleave RNA in the ribosome, thus inactivating the ribosome and ultimately killing the cell (Kao & Davies, 1995). The RNA cleavage is carried out by a mechanism quite similar to that found in the T_1 ribonucleases (Campos-Olivas *et al.*, 1996). The

Table 4. Results of application of threading and the T_1 ribonuclease FFF to nine ribonuclease sequences from various organisms and three ribotoxin sequences from the *Aspergillus* fungi family

Sequence	Signif.	Active-site Res. (H H E Y Phob R)
RNT1_ASPOR (9rnt)	132104. 645277, 4334	40 92 58 38 100 77
RNF1_FUSMO (1fus)	2358, 6673, 428	30 90 57 37 98 75
RNMS_ASPSA (1rms)	12586, 38926, 1244	39 91 57 37 99 76
RNU2_USTSP (1rtu)	122, 332, 61.0	41 101 62 39 110 85
RNC2_ASPCL	18638, 88768, 1321	40 92 58 38 100 77
RNPB_PENBR	6650, 26930, 810	38 90 56 36 98 75
RNPC_PENCH	5977, 21676, 764	38 90 56 36 98 75
RNN1_NEUCR	11840, 47527, 946	40 92 58 38 100 77
RNU1_USTSP	677, 1690, 132	37 92 57 35 100 76
RNAS_ASPGI (α-sarcin)	26.4, 72.5, 14.8	77 164 123 75 172 148
RNCL_ASPCL (clavin)	26.1, 82.2, 13.3	77 164 123 75 172 148
RNMG_ASPRE (restrictocin)	21.1, 91.2, 15.8	76 163 122 74 171 147

Signif. is the significance score of the alignment. The first entry of the group of three is the sequence-based scoring method; the second entry is the sequence-structure scoring method; the third is the structure-structure scoring method. All sequences aligned with the 9rnt structure. Active-site res. are the residues in the threading model that are identified by the RNA hydrolysis T_1 ribonuclease FFF as being active-site residues. The first group (H H E) is the triad involved in the nucleophilic catalysis; the second group (Y Phob R) is involved in transition state stabilization (see the text for more details). For the ribotoxins, the residue numbering includes the leader sequence, as shown in Figure 7A.

three selected ribotoxins, α-sarcin (RNAS_ASPGI), clavin (RNCL_ASPCL) and restrictocin (mitogillin, RNMG_ASPRE), can be aligned with the T_1 ribonucleases by multiple sequence alignment algorithms (Figure 7A), but the degree of sequence identity between the ribotoxins and the T_1 ribonucleases is quite low (less than 35% pairwise sequence identity). Furthermore, a Blast (Altschul *et al.*, 1990) search of SwissProt (Bairoch & Apweiler, 1996) using the sequence of 9rnt as the search sequence does not yield any of these ribotoxin sequences. The structure of restrictocin (Yang & Moffat, 1996) was solved and recently released; that of α-sarcin (Campos-Olivas *et al.*, 1996) was solved, but has not yet been released to the public databases.

The three ribotoxin sequences, including their signal sequences, were threaded through 301 nonhomologous protein structures (Fischer *et al.*, 1996). As with the T_1 ribonucleases, each ribotoxin sequence aligned with 9rnt as the highest-scoring sequence by all three scoring methods, although the alignment scores were much lower than those for the T_1 ribonucleases themselves (Table 4). Nine models (three sequences times three scoring methods) were built based on the sequence-to-structure alignments produced by the threading program. All nine models contained both the nucleophilic and the transition state stabilization triads and were recognized by the T_1 ribonuclease FFF. The identified active-site residues are presented in Table 4. This result again demonstrates that models of distantly related proteins can be built on sequence-to-structure alignments produced by a threading algorithm. Active sites within these low-to-moderate resolution models can be recognized by the FFF.

None of the T_1 ribonuclease sequences has yet been folded by the MONSSTER algorithm; however, as a control, this FFF was tested on a database of correctly and incorrectly folded structures produced by MONSSTER and used to test the glutaredoxin/thioredoxin FFF. None of these structures was found to contain the T_1 RNase active site, even when the allowed variance in the distance was uniformly increased by ± 0.5 Å over those variances shown in Table 2.

Discussion

With the advent of the genome sequencing projects, the number of known protein sequences is exponentially increasing; however, the sequence of a protein is virtually useless without some knowledge of both its structure and its function. The most common methods for predicting protein function from sequence are to look for homologous proteins in the sequence databases by standard sequence alignment protocols, or to look for local sequence signatures that match those found in the appropriate functional databases such as Prosite, Blocks and Prints.

Here, we have demonstrated the utility of a new method for predicting protein function based on the three-dimensional structure of the active site. This method is based on the sequence-to-structure-to-function paradigm, because the structure of the protein is first predicted from its sequence, then the active site of the protein is identified in the predicted model. We have shown here that active-site descriptors, termed FFFs, work to identify the active-site residues both in high-resolution (exact) and low-to-moderate resolution (inexact or predicted) protein structures.

Advantages of using geometric descriptors to identify protein active sites

Because the method is based on three-dimensional structures of protein active sites, it has the

following advantages (each is discussed in further detail in the following paragraphs): (1) it is applicable even when the degree of sequence identity between two proteins is not significant; (2) it can, in principle, treat the case of proteins having two different global folds, but similar sites and associated function; (3) it distinguishes between proteins with similar folds (topological cousins) and those that belong to a given functional family; and (4) in addition to assigning a given protein to a functional family, the method produces a map or model of the protein active site.

The examples presented in the Introduction suggest that functionally important residues are often non-local in sequence and important functional relationships cannot always be extracted from standard sequence comparisons. As the sequence and structural databases grow ever larger, examples such as these will become increasingly more common. Motif databases such as Prosite (Bairoch *et al.*, 1995), Blocks (Henikoff & Henikoff, 1991) and Prints (Attwood & Beck, 1994; Attwood *et al.*, 1994, 1997), while very powerful, are limited in scope because they are restricted to one-dimensional sequence information. FFFs, because they are based on the three-dimensional structure of the active site, should be able to identify the similar function in these families, as was shown here for the oxidoreductase activity of the glutaredoxin/thioredoxin family (Table 3) and the RNA hydrolytic activity of the T_1 ribonucleases (Table 4). We have shown several cases where the FFF was able to identify an active site when the sequence identity between the two sequences was in the twilight zone. We have predicted the active site and associated activity in one case where the sequence identity is insignificant and the function is not identified by Prosite, Prints or Blocks. Finally, we predicted the activity in two cases where neither BLAST nor the motif databases predict the glutaredoxin/thioredoxin active site (Table 3).

In the extreme case, however, two proteins might have similar active sites even though their tertiary structures are completely different, as found for the mammalian and bacterial serine proteases (Branden & Tooze, 1991). In such a case, sequence alignment or local sequence signatures would be unable to recognize the functional similarity. The method presented here has the advantage that it should be able to recognize the active-site similarity in such cases. Such cases will be examined in the future.

The third advantage of the method is that it can distinguish between topological cousins and proteins having similar function. The number of known folds is not increasing as quickly as the number of solved structures. For instance, the structural family of the α/β barrel proteins is quite large; however, the members of this family can have quite disparate functions. This has led some researchers to suggest that there are a limited number of protein folds (Godzik, 1997; Holm & Sander, 1997a; Orengo *et al.*, 1994; Wang, 1996), a statement that bodes well for prediction of protein structure by threading or inverse folding-based approaches. However, while this observation might increase the chances of predicting a structure *via* threading, it decreases our ability to predict the function of that same protein *via* threading. If many different proteins with differing functions fold into similar structures, simple structure prediction will tell us nothing whatsoever about the function of the protein. The results for YEO4_YEAST and YPRO82c demonstrate that the FFF can provide a method for automatic distinction between functionally related and topological cousins. Thus, a library of FFFs would greatly expand the utility of current threading algorithms and allow us to predict protein function, as well as structure, *via* threading approaches.

Use of the sequence-to-structure-to-function paradigm for prediction of protein function confers one further set of advantages: the predicted structure produces a model of the protein and the FFF identifies the exact location of the active site in both the sequence and its predicted structure. In contrast, standard sequence analysis methods, being inherently one-dimensional, do not automatically provide a model of the active site. Once the protein's function is predicted by sequence homology, a model of the active site can be built by homology modeling based on the sequence alignment, provided that the structure of a related protein is known. In the FFF paradigm, the procedure is inverted: first, a model structure is built, then it is scanned for possible location of the active site. Once the location of the active site is identified, the active site and the model can be further studied for possible similarities to or differences from other active sites of the same protein family. For example, once a predicted structure is identified as having the redoxin disulfide oxidoreductase activity, the structure can be analyzed in more detail to see if it belongs to the thioredoxin or the glutaredoxin subfamilies or to another, as yet unidentified, subfamily. Examples of active-site residues predicted by the FFFs described here are presented in Tables 3 and 4. Finally, the sequence whose function is newly assigned can now be used to scan the sequence databases for other homologous sequences that are compatible with the predicted function.

Disadvantages of using geometric descriptors for identifying protein active sites

The FFF approach suffers from several disadvantages. First, a structure of the protein must be available. This is not as great a disadvantage as it might seem, because we have shown here that low-resolution models produced by current prediction algorithms are still useful for active-site prediction using this method. Protein structure prediction tools and algorithms will only improve, but even at this stage, useful models can still be produced. A second disadvantage is that the resulting model might actually be incorrect, i.e. it

is misfolded, either globally or locally. Such an incorrect structure could cause misidentification of an active site, with either a false positive or a false negative result, depending on the particular case. As the method is further tested, such situations will undoubtedly be observed. The final and major disadvantage is that the active site responsible for the protein's function must have been previously observed and studied. Otherwise, using the current method, it is not possible to build a FFF. However, we will start by building a library of FFFs with the many active sites that have been well studied and whose structures are known. Even if our approach is limited to previously identified active sites, the ability to predict proteins that have these active sites will still be very useful and will extend the limits of function prediction from sequence much further into the twilight zone.

Conclusions and Future Directions

Here, we have shown that simple, relaxed geometric and conformational descriptors (FFFs) of the active sites of proteins are sufficient to select proteins containing specific activities from a large set of high-resolution models. We have shown that the two developed FFFs are specific for low-resolution models created either by *ab initio* folding algorithms or by threading algorithms. Finally, we have presented an example where the FFF predicts the function of two proteins from the yeast genome whose structures were predicted by a threading algorithm and whose function could not be identified by local signature databases or by Blast. This work increases the utility of the genomic sequence databases and demonstrates that predicted models, even those at low resolution, can be used for protein function prediction. Thus, it paves the way for the functional screening of the genomic databases based on the sequence-to-structure-to-function paradigm, provided that the active-site geometry and conformation is found in a previously solved structure.

In the future, we plan to expand the library of FFFs to include many more active sites, focusing on those activities that are not easily identifiable by local sequence motifs. We will include the peptide hydrolytic active site of the serine proteases in the expanded FFF library to show that FFFs can identify similar active sites even when the global fold is completely different. The threading results presented here suggest that screening of complete genomes for function might be feasible, and this, too, will be attempted. Analysis of complete genomes will provide more detailed analysis on the success and failure rate of this method.

Methods

Description of how to build an FFF

The FFFs are built from the three-dimensional structural arrangements of functionally important residues on the basis of the biochemistry of the known function. These geometric descriptors should be inherently more exact than local sequence signatures, because they encode structural as well as minimal sequence information and, thus, they will be more descriptive of the actual chemistry involved in the protein function. A general outline of how to build a FFF is shown in Figure 3.

The first step is to perform a literature search to gather biochemical evidence about which residues are functionally important. Next, a series of functionally related proteins with known structures are selected. These putative functionally important residues are superimposed in space, and their relative geometries (distances, angles) between α-carbon atoms and side-chain centers of mass are recorded. Common secondary structures are identified, if there is evidence in the literature for the importance of such conformations. Structural superposition and multiple sequence alignment can help identify other residues that might be important, but these should be used only if experimental evidence suggests a functional significance. The procedure is iterative. After identification of conserved residues, another literature search can be done to analyze the relative functional importance of these conserved residues and structures. We aim to use only those residues shown to be functionally important or conserved across a large set of proteins exhibiting the activity of interest.

Once a set of geometric and conformational constraints for a specific function has been identified, they are implemented in the form of a computer algorithm. The program searches experimentally determined protein structures from the protein structural databank (Abola *et al.*, 1987) for sets of residues that satisfy the specified constraints. The constraints are implemented stepwise, so that structures that are eliminated by each criterion can be evaluated at each step along the way. If the constraint set misses any proteins known to exhibit the function under investigation, the structure of the missed protein is analyzed and the FFF modified. If the FFF selects proteins that are not known to display the function, then the structure of these ''false positive'' examples is compared to the known functional sites. Again, the FFF is modified to eliminate the false positives, although some false positives could prove to be interesting if they identify a previously unrecognized activity in a protein.

At this stage, a tentative FFF is generated that can be applied to structures of varying quality (Figure 3). While the FFF is initially tuned to high-resolution structures, it might be loosened to accommodate ambiguities inherent in lower-resolution models. Ideally, such fuzziness should not degrade the performance on high-resolution structures. Thus, the extent of fuzziness is ascertained by the performance on exact (i.e. a set of high-resolution) structures and on low-resolution models of known structure.

Using this method, FFFs were created for the disulfide oxidoreductase activity of the glutaredoxin/thioredoxin family and the RNA hydrolytic activity of the T_1 ribonuclease family. The information from the literature about the enzymatic reaction mechanism used to create these FFFs is described in Results. For the disulfide oxidoreductase FFF, 1aaz chains A and B (Eklund *et al.*, 1992), 1dsb chain A (Martin *et al.*, 1993) and 4trx (Forman-Kay *et al.*, 1990) were used to define the active-site geometric information. For the T_1 ribonuclease FFF, 1rtu (Noguchi *et al.*, 1995), 1fus (Vassylyev *et al.*, 1993) and 1rms (Nonaka *et al.*, 1993) were used to define the active site.

The final α-carbon to α-carbon distances and their allowed variances that describe the final FFFs are compiled in Table 2. Most often, these distances are similar to the average distances for proteins used to define the FFF and the allowed variance correlates with the standard deviation (see Table 2 for examples). The FFFs themselves and their application to actual and predicted structures are described in more detail in Results.

Description of the threading or inverse folding algorithm

In an inverse folding approach, one ''threads'' a probe sequence through different template structures and attempts to find the most compatible structure for a given sequence. The threading program used here is that created and distributed by Godzik and co-workers (Jaroszewski *et al*., 1998). Briefly, the sequence-to-structure alignments are performed by a ''local-global'' version of the Smith-Waterman algorithm (Waterman, 1995). The alignments are then ranked by three different scoring methods (Jaroszewski *et al*., 1998). The first, SQ, is based on a sequence-sequence type of scoring. In this sequence-based method, the Gonnet mutation matrix was used to optimize gap penalties, as described by Vogt and Argos (Vogt *et al*., 1995). The second method, BR, is a sequence-structure scoring method that is based on the pseudo-energy from the probe sequence ''mounted'' in the structural environment in the template structure. The pseudo-energy term reflects the statistical propensity of successive amino acid pairs (from the probe sequence) to be found in particular secondary structures within the template structure. The third method, TT, is a structure-structure scoring method, whereby information from the known template structure is compared to the predicted secondary structure of the probe sequence. The secondary structure prediction scheme for the probe sequence was the nearest-neighbor algorithm (L. Rychlewski & A. Godzik, unpublished). The version used here achieves an average three-state prediction accuracy of 74%.

Once we have computed scores for the sequence-to-structure alignments, the statistical significance of each score must be determined. To determine this significance, the distribution of scores is fit to an extreme value distribution and the raw score is compared to the chance of obtaining the same score when comparing two unrelated sequences, as described by Godzik and co-workers (Jaroszewski *et al*., 1998). Tables 3 and 4 report the significance score of the top sequence-to-structure alignments, rather than the raw score.

Once the alignment of the probe sequence-to-template structure has been determined, a three-dimensional model must be built. Scripts utilizing the automatic modeling tools provided by Modeller4 (Sali & Blundell, 1993) were developed (L. Jaroszewski, K. Pawlowski & A. Godzik, unpublished). These scripts automatically produce all-atom coordinate files for the three-dimensional model built from the sequence-to-structure alignment provided by the threading algorithm. The FFF was applied directly to these structures without any further enhancement, energy calculations or molecular mechanics simulations of the model.

Description of MONSSTER, the *ab initio* folding algorithm

Some predicted structures were produced using a method for the *ab initio* prediction of protein structures at low resolution (Ortiz *et al*., 1998; Skolnick *et al*., 1997). Predicted structures used for the FFF analysis were taken directly from the set of correctly and incorrectly folded proteins produced by this procedure. Briefly, the procedure can be divided into two parts: restraint derivation using information extracted from multiple sequence alignment and structure assembly/refinement using an improved version of the MONSSTER algorithm, which uses a high coordination lattice-based C^{α} representation for the folding of proteins (Skolnick *et al*., 1997), modified to incorporate the expected accuracy and precision of the predicted tertiary restraints (Ortiz *et al*., 1998).

For each protein sequence, 10 to 40 independent simulated annealing simulations from a fully extended initial conformation were carried out (assembly runs). Structures were then clustered and all low-energy structures were subjected to low-temperature, isothermal refinement. The predicted fold is that of lowest average energy. The FFFs were tested on a series of correctly and incorrectly folded structures produced during the assembly and isothermal runs for proteins 1ego (Xia *et al*., 1992), 1poh (Jia *et al*., 1993), 1ubq (Vijay-Kumar *et al*., 1987) and 1cis (Osmark *et al*., 1993).

Acknowledgments

The authors thank A. Godzik and A. Ortiz for stimulating discussions. This work was supported, in part, by a grant from Johnson and Johnson. J.S.F. gratefully acknowledges the hospitality of The Scripps Research Institute during her sabbatical leave.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Protein Data Bank in Crystallographic Databases–Information Content, Software Systems, Scientific Application* (Allen, F. H., Bergerhoff, G. & Sievers, R., eds), Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Aszodi, A., Gradwell, M. J. & Taylor, W. R. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **251**, 308–326.

Attwood, T. K. & Beck, M. E. (1994). PRINTS–a protein motif fingerprint database. *Protein Eng.* **7**, 841–848.

Attwood, T. K., Beck, M. E., Bleasby, A. J. & Parry-Smith, D. J. (1994). PRINTS–a database of protein motif fingerprints. *Nucl. Acids Res.* **22**, 3590–3596.

Attwood, T. K., Beck, M. E., Bleasby, A. J., Degtyarenki, K., Michie, A. D. & Parry-Smith, D. J. (1997). Novel developments with the PRINTS protein fingerprint database. *Nucl. Acids Res.* **25**, 212–216.

Bairoch, A. & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucl. Acids Res.* **24**, 21–25.

Bairoch, A., Bucher, P. & Hofmann, K. (1995). The PROSITE database, its status in 1995. *Nucl. Acids Res.* **24**, 189–196.

Biniszkiewicz, D., Cesnaviciene, E. & Shub, D. A. (1994). Self-splicing group I intron in cyanobacterial initiator methionine tRNA: evidence for lateral

transfer of introns in bacteria. *EMBO J.* **13**, 4629–4635.

Blattner, F. R., Plunkett, I., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1453–1462.

Bleasby, A. J., Adrigg, D. & Attwood, T. K. (1994). OWL–a non-redundant, composite protein sequence database. *Nucl. Acids Res.* **22**, 3574–3577.

Branden, C. & Tooze, J. (1991). *Introduction to Protein Structure*, Garland Publishing, Inc., New York.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A. & Scott, J. L. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science,* **273**, 1058–1073.

Bushweller, J. H., Aslund, F., Wuthrich, K. & Holmgren, A. (1992). Structural and functional characterization of the mutant *Escherichia coli* glutaredoxin (C14–S) and its mixed disulfide with glutathione. *Biochemistry,* **13**, 9288–9293.

Campos-Olivas, R., Bruix, M., Santoro, J., Pozo, A. M. D., Lacadena, J., Gavilanes, J. G. & Rico, M. (1996). Structural basis for the catalytic mechanism and substrate specificity of the ribonuclease alpha-sarcin. *FEBS Letters,* **399**, 163–165.

Collins, F. & Galas, D. (1993). A new five-year plan for the U.S. Human Genome Project. *Science,* **262**, 43–46.

Day, C. L., Anderson, B. F., Tweedie, J. W. & Baker, E. N. (1993). Structure of the recombinant N-terminal lobe of human lactoferrin at 2.0 Å resolution. *J. Mol. Biol.* **232**, 1084–1100.

Deisenhofer, J., Epp, O., Sinning, I. & Michel, H. (1995). Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis*. *J. Mol. Biol.* **246**, 429–457.

Delseny, M., Cooke, R., Raynal, M. & Grellet, F. (1997). The *Arabidopsis thaliana* cDNA sequencing projects. *FEBS Letters,* **405**, 129–132.

Dujon, B. (1996). The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270.

Dyson, H. J., Jeng, M. F., Tennant, L. L., Slaby, I., Lindell, M., Dui, D. S., Kuprin, S. & Holmgren, A. (1997). Effects of buried charged groups on cysteine thiol ionization and reactivity in *Escherichia coli* thioredoxin: structural and functional characterization of mutants of Asp16 and Lys57. *Biochemistry,* **36**, 2622–2636.

Eklund, H., Ingelman, M., Soderberg, B. O., Uhlin, T., Nordlund, P., Nikkola, M., Sonnerstam, U., Joelson, T. & Petratos, K. (1992). Structure of oxidized bacteriophage T4 glutaredoxin (thioredoxin). Refinement of native and mutant proteins. *J. Mol. Biol.* **228**, 596–618.

Fischer, D., Rice, D., Bowie, J. U. & Eisenberg, D. (1996). Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.* **10**, 126–136.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Field, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, J., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269**, 496–512.

Forman-Kay, J. D., Clore, G. M. & Gronenborn, A. M. (1990). Studies on the solution conformation of human thioredoxin using heteronuclear ^{15}N-^{1}H nuclear magnetic resonance spectroscopy. *Biochemistry,* **29**, 1566–1572.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, D.-C., Lucier, T. S., Peterson, S. N., Smith, H. O. C. A., Hutchison, I. & Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science,* **270**, 397–403.

Friesner, R. & Gunn, J. R. (1996). Computational studies of protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 315–342.

Fromont-Racine, M., Rain, J. C. & Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genet.* **16**, 277–282.

Gibbs, R. (1995). Pressing ahead with human genome sequencing. *Nature Genet.* **11**, 121–125.

Godzik, A. (1997). Counting and classifying possible protein folds. *Trends Biotech.* **15**, 147–151.

Goldberg, J., Huang, H. B., Kwon, Y. G., Greengard, P., Nairn, A. C. & Kuriyan, J. (1995). Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature,* **376**, 745–753.

Hellinga, H. W. & Richards, F. M. (1991). Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.* **222**, 763–785.

Hellinga, H. W., Caradonna, J. P. & Richards, F. M. (1991). Construction of new ligand binding sites in proteins of known structure. II. Grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin. *J. Mol. Biol.* **222**, 787–803.

Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl Acids Res.* **19**, 6565–6572.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. & Herrmann, R. (1997). Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucl. Acids Res.* **25**, 701–712.

Hobohm, U. & Sander, C. (1995). A sequence property approach to searching protein databases. *J. Mol. Biol.* **251**, 390–399.

Holm, L. & Sander, C. (1997a). Dali/FSSP classification of three-dimensional folds. *Nucl. Acids Res.* **25**, 231–234.

Holm, L. & Sander, C. (1997b). An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins: Struct. Funct. Genet.* **28**, 72–82.

Holmgren, A. & Aslund, F. (1995). Glutaredoxin. *Methods Enzymol.* **252**, 283–292.

Holmgren, A. & Bjornstedt, M. (1995). Thioredoxin and thioredoxin reductase. *Methods Enzymol.* **252**, 199–208.

Ito, T. & Sakaki, Y. (1996). Toward genome-wide scanning of gene expression: a functional aspect of the genome project. *Essays Biochem.* **31**, 11–21.

Jaroszewski, L., Rychlewski, L., Zhang, B. & Godzik, A. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Proteins Sci.* **7**, 1431–1440.

Jia, Z., Quail, J. W., Waygood, E. B. & Delbaere, L. T. (1993). The 2.0 angstrom resolution structure of *Escherichia coli* histidine-containing phosphocarrier protein HPr. A redetermination. *J. Biol. Chem.* **268**, 22490–22501.

Kao, R. & Davis, J. (1995). Fungal ribotoxins: a family of naturally engineered targeted toxins?. *Biochem. Cell. Biol.* **73**, 1151–1159.

Katti, S. K., LeMaster, D. M. & Eklund, H. (1990). Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* **212**, 167–184.

Katti, S. K., Robbins, A. H., Yang, Y. & Wells, W. W. (1995). Crystal structure of thioltransferase at 2.2 Å resolution. *Protein Sci.* **4**, 1998–2005.

Kemmink, J., Darby, N. J., Dijkstra, K., Scheek, R. M. & Creighton, T. E. (1995). Nuclear magnetic resonance characterization of the N-terminal thioredoxin-like domain of protein disulfide isomerase. *Protein Sci.* **4**, 2587–2593.

Kemmink, J., Darby, N. J., Dijkstra, K., Nilges, M. & Creighton, T. E. (1997). The folding catalyst protein disulfide isomerase is constructed of active and inactive thioredoxin modules. *Curr. Biol.* **7**, 239–245.

Kortemme, T. & Creighton, T. E. (1995). Ionisation of cysteine residues at the termini of model alpha-helical peptides. Relevance to unusual thiol pK_a values in proteins of the thioredoxin family. *J. Mol. Biol.* **253**, 799–812.

Kortemme, T. & Creighton, T. E. (1996). Electrostatic interactions in the active site of the N-terminal thioredoxin-like domain of protein disulfide isomerase. *Biochemistry,* **35**, 14503–14511.

Martin, J. L., Bardwell, J. C. & Kuriyan, J. (1993). Crystal structure of the DsbA protein required for disulphide bond formation *in vivo*. *Nature,* **365**, 464–468.

Martinez-Oyanedel, J., Choe, H.-W., Heinemann, U. & Saenger, W. (1991). Ribonuclease T_1 with free recognition and catalytic site: crystal structure at 1.5 Å resolution. *J. Mol. Biol.* **22**, 335–352.

Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F. & Zollner, A. (1997). Overview of the yeast genome. *Nature,* **387**, 1–65.

Mumenthaler, C. & Braun, W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.* **4**, 863–871.

Noguchi, S., Satow, Y., Uchida, T., Sasaki, C. & Matsuzaki, T. (1995). Crystal structure of *Ustilago sphaerogena* ribonuclease U2 at 1.8 Å resolution. *Biochemistry,* **34**, 15583–15591.

Nonaka, T., Mitsui, Y., Uesugi, S., Ikehara, M., Irie, M. & Nakamura, K. T. (1993). Crystal structure of ribonuclease Ms (as a ribonuclease T1 homologue) complexed with a guanylyl-3,5-cytidine analogue. *Biochemistry*, **32**, 11825–11837.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631–634.

Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998). Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* **277**, 419–448.

Osmark, P., Sorensen, P. & Poulsen, F. M. (1993). Context dependence of protein secondary structure formation: the three-dimensional structure and stability of a hybrid between chymotrypsin inhibitor 2 and helix E from subtilisin Carlsberg. *Biochemistry,* **32**, 11007–11014.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227–258.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci, USA,* **85**, 2444–2448.

Rohrer, J. & Rawlings, D. E. (1992). Sequence analysis and characterization of the mobilization region of a broad-host-range, pTF-FC2, isolated from *Thiobacillus ferrooxidans*. *J. Bacteriol.* **174**, 6230–6237.

Sali, A. & Blundell, T. L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.

Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997). MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217–241.

Smith-Brown, M. J., Kominos, D. & Levy, R. M. (1993). Global folding of proteins using a limited number of distance constraints. *Protein Eng.* **6**, 605–614.

Soman, J., Iismaa, S. & Stout, C. D. (1991). Crystallographic analysis of two site-directed mutants of *Azotobacter vinelandii* ferredoxin. *J. Biol. Chem.* **266**, 21558–21562.

Srinivasan, R. & Rose, G. D. (1995). LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins: Struct. Funct. Genet.* **22**, 81–99.

Steyaert, J. (1997). A decade of protein engineering on ribonuclease T1–atomic dissection of the enzyme-substrate interactions. *Eur. J. Biochem.* **247**, 1–11.

Sturrock, S. S. & Collins, J. F. (1993). MPsrch version 1.3. In *Biocomputing Research Unit*, University of Edinburgh, UK.

Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73.

Tauer, A. & Benner, S. A. (1997). The B_{12}-dependent ribonucleotide reductase from the archaebacterium *Thermoplasma acidophila*: an evolutionary solution to the ribonucleotide reductase conundrum. *Proc. Natl Acad. Sci. USA,* **94**, 53–58.

Vassylyev, D. G., Katayanagi, K., Ishikawa, K., Tsujimoto-Hirano, M., Danno, M., Pahler, A., Matsumoto, O., Matsushima, M., Yoshida, H. & Morikawa, K. (1993). Crystal structures of ribonuclease F1 of *Fusarium moniliforme* in its free form and in complex with 2′ GMP. *J. Mol. Biol.* **230**, 979–996.

Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194**, 531–544.

Vogt, G., Etzold, T. & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* **249**, 816–831.

Wang, Z. X. (1996). How many fold types of proteins are there in nature?. *Proteins: Struct. Funct. Genet.* **26**, 186–191.

Waterman, M. S. (1995). *Introduction to Computational Biology. Maps, Sequences, and Genomes*, Chapman & Hall, London.

Xia, T.-H., Bushweller, J. H., Sodano, P., Billeter, M., Bjornberg, O., Holmgren, A. & Wuthrich, K. (1992). NMR structure of oxidized *Escherichia coli* glutaredoxin: comparison with reduced *E. coli* glutaredoxin and functionally related proteins. *Protein Sci.* **1**, 310–321.

Yang, X. & Moffat, K. (1996). Insights into specificity of cleavage and mechanism of cell entry from the crystal structure of the highly specific *Aspergillus* ribotoxin, restrictocin. *Structure*, **4**, 837–852.

Yang, Y. F. & Wells, W. W. (1991a). Identification and characterization of the functional amino acids at the active center of pig liver thioltransferase by site-directed mutagenesis. *J. Biol. Chem.* **266**, 12759–12765.

Yang, Y. F. & Wells, W. W. (1991b). Catalytic mechanism of thioltransferase. *J. Biol. Chem.* **266**, 12766–12771.

Edited by F. Cohen

(Received 18 November 1997; received in revised form 8 June 1998; accepted 8 June 1998)