Protein Science (1999), *8*:1104–1115. Cambridge University Press. Printed in the USA.

From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions

BAOHONG ZHANG,[1] LESZEK RYCHLEWSKI,[2] KRZYSZTOF PAWŁOWSKI,[3]
JACQUELYN S. FETROW, JEFFREY SKOLNICK, AND ADAM GODZIK[3]
The Scripps Research Institute, La Jolla, California 92037

(Received October 26, 1998; Accepted February 12, 1999)

Abstract

A database of functional sites for proteins with known structures, SITE, is constructed and used in conjunction with a simple pattern matching program SiteMatch to evaluate possible function conservation in a recently constructed database of fold predictions for *Escherichia coli* proteins (Rychlewski L et al., 1999, *Protein Sci 8*:614–624). In this and other prediction databases, fold predictions are based on algorithms that can recognize weak sequence similarities and putatively assign new proteins into already characterized protein families. It is not clear whether such sequence similarities arise from distant homologies or general similarity of physicochemical features along the sequence. Leaving aside the important question of nature of relations within fold superfamilies, it is possible to assess possible function conservation by looking at the pattern of conservation of crucial functional residues. SITE consists of a multilevel function description based on structure annotations and structure analyses. In particular, active site residues, ligand binding residues, and patterns of hydrophobic residues on the protein surface are used to describe different functional features. SiteMatch, a simple pattern matching program, is designed to check the conservation of residues involved in protein activity in alignments generated by any alignment method. Here, this procedure is used to study conservation of functional features in alignments between protein sequences from the *E. coli* genome and their optimal structural templates. The optimal templates were identified and alignments taken from the database of genomic structural predictions was described in a previous publication (Rychlewski L et al., 1999, *Protein Sci 8*:614–624). An automated assessment of function conservation is used to analyze the relation between fold and function similarity for a large number of fold predictions. For instance, it is shown that identifying low significance predictions with a high level of functional residue conservations can be used to extend the prediction sensitivity for fold prediction methods. Over 100 new fold/function predictions in this class were obtained in the *E. coli* genome. At the same time, about 30% of our previous fold predictions are not confirmed as function predictions, further highlighting the problem of function divergence in fold superfamilies.

Keywords: fold assignments; function predictions; genome analysis

The prediction of protein folds and functions from sequence is the "Holy Grail" of molecular biology. With improving sequencing methods, the number of known protein sequences has increased over 10-fold in the last two years and is expected to grow even faster in the next several years. The experimental characterization of new proteins is also improving, but at a much slower rate. Consequently, computer analysis of new sequences, particularly aiming at recognition of similarity to the already characterized protein families, has become a primary tool for analysis of new sequences. For instance, most newly sequenced genomes were first analyzed by tools such as BLAST (Altschul et al., 1990) or FASTA (Pearson & Miller, 1992), and the results of this analysis were the primary source of most annotations present in sequence databases.

This type of analysis is based on the "similar sequence–similar structure–similar function" rule. Most often, this rule is applied to closely homologous proteins where sequence similarity is easily recognizable. For such proteins, both their structures and their functions are usually similar. With accumulating experimental data about thousands of proteins, many examples of proteins with similar folds but no apparent sequence similarity were discovered.

Reprint requests to: Adam Godzik, The Burnham Institute, 10901 North Torrey Pines Rd., La Jolla, California 92037; e-mail: adam@burnham-inst.org.

[1] Present address: Stanford University, Department of Mathematics, Bldg. 380, Stanford, California 94305.

[2] Present address: San Diego Supercomputer Center, UCSD, 9500 Gilman Drive, La Jolla, California 92093.

[3] Present address: The Burnham Institute, 10901 N. Torrey Pines Rd., La Jolla, California 92037.

Some level of functional similarity sometimes, but not always, accompanies such structural similarity. It is not clear what the evolutionary relationship between such proteins is. Arguments for distant evolutionary relationship, convergent evolution, and random similarity are often made in the same cases (Murzin, 1998). However, from the point of view of function prediction, we are faced with similar question in both "distant homology" and "random similarity" (or "convergent evolution") scenarios of function prediction.

Many new sequence analysis methods developed in the last few years attempt to recognize such proteins by extending the notion of sequence similarity beyond a simple mutation matrix. This is done by including additional information about one or both of the proteins being compared. Two classes of methods seem to exemplify two possible solutions for the "similar fold–not similar sequence" puzzle. For instance, in the "profile" methods (Gribskov et al., 1987; Bork & Gibson, 1996; Altschul et al., 1997; Rychlewski et al., 1998), sequence information is enhanced by a mutation pattern on a given position along the sequence. In the "threading" methods (Bowie et al., 1991; Godzik et al., 1992; Jones et al., 1992; Russell et al., 1996; Jaroszewski et al., 1998), sequence information is enhanced or replaced by residue interaction preferences, thus aiming at identifying the "structural" signature of the sequence and recognition of the structural similarity even in the absence of homology.

The "profile" methods have a wider application, because they could recognize similarities between proteins, for which none of the structures is known. However, "profile" methods can be applied to the problem of structure prediction by limiting the database of proteins used for comparison to proteins with known structures. Surprisingly, despite their different points of origin, both profile and threading methods seem to give similar results, at least in limited tests (Rychlewski et al., 1998; also see Methods). This makes the "profile" methods practically, if not logically, similar to the threading methods. Therefore, in this, as well as in previous papers (Pawlowski et al., 1999; Rychlewski et al., 1998), we are using a sequence based method for fold assignments.

Does it make sense to divide the problem of identification of a new protein into two subproblems: an unrestricted search for any similar proteins and a restricted search for similar protein with known three-dimensional (3D) structure? Here and in the previous papers we argue that it does. In particular, we argue that knowledge of the full 3D structure allows much deeper understanding of the protein function. This understanding gives us the opportunity for additional verification of both fold and function prediction, and in many cases, some of which were discussed in detail in the previous publications, allows to make additional predictions as to the molecular level function for new proteins. In this contribution, we attempt to automate some of these "next level" analysis of the fold prediction.

Function predictions are often made implicitly after some level of sequence similarity is detected between two proteins. Many newly sequenced proteins are annotated as "putative homologues" of some well-characterized proteins, with an implicit assumption that their function must be similar to that of its putative homologue. However, the function prediction could be wrong, even if the two proteins are homologous, because of the divergence of functions in homologous proteins. It could also be right, even if the two proteins are not homologous, because the sequence similarity could be a result of convergent evolution. At the same time, the functional prediction could be easily verified by checking the level of conservation of functionally important residues. Such verification is intuitively obvious, and on numerous occasions was done by various authors in the analysis of specific protein families.

With improvement in fold assignment algorithms, whether by threading or profile methods, several groups attempted genome scale analyses for microbial genomes: *Mycoplasma genitalium* (Fischer & Eisenberg, 1997; Huynen et al., 1998; Jones, 1998; Rychlewski et al., 1998), *Escherichia coli* (Rychlewski et al., 1999) and *Helicobacter pylori* (Pawlowski et al., 1999). In each of these papers, several examples of fold predictions in each genome were analyzed manually for possible function conservation. In addition, an algorithm for functional analysis of 3D protein models (Fetrow & Skolnick, 1998) was used to screen fold predictions obtained from threading (Jaroszewski et al., 1998) for proteins that may possess glutaredoxin/thioredoxin activity for proteins from *E. coli* (Fetrow et al., 1998). Here, both efforts are combined in an attempt to generate a wide survey of the possibility of following fold prediction with function prediction. In particular, the previous fold assignments for proteins from the *E. coli* genome (Rychlewski et al., 1999) are now complemented with an automated function assessment based on conservation of the functional site residues.

The paper is constructed as follows. In Results, the SiteMatch program is used to analyze function conservation in the database of fold predictions for proteins from the *E. coli* genome (Rychlewski et al., 1999). The program uses SITE, a database of multilevel protein function description, built from records preset in Protein Data Bank (PDB) files as well as from direct analysis of structure files. At present, the database contains information about active site residues, ligand binding sites, and potential protein–protein interfaces. The main purpose of the SiteMatch program is to annotate the alignments and thus make the prediction analysis easier and faster. At the same time, some observations about a general picture of function conservation among proteins predicted to belong to already known fold superfamilies can be made and are presented in Results. Examples of closely homologous proteins that have lost their activity highlight the differences between the homology and function predictions. Next, examples of predictions that were below the previously used significance thresholds and were therefore discarded in the previous analysis (Rychlewski et al., 1999) are discussed in the second part of Results. These examples highlight the use of fold prediction (with or without homology) as a first step in function prediction. Finally, the analysis of hydrophobic pattern conservation is used to predict the oligomerization state of new proteins. "SITE," a database of multilevel protein structure functional annotations, and "SiteMatch," a pattern-matching program to analyze the conservation of specific functional residues in alignments, are described in Methods. Other algorithms used in the paper are also described. Throughout the text, the first part of a SwissProt (Bairoch & Apweiler, 1999) name is used to identify proteins; thus, for example, the name YRAR refers to YRAR_ECOLI SwissProt entry.

Results

Function prediction vs. fold prediction

Fold predictions for the set of 4,287 protein sequences from the *E. coli* genome were adopted from previous work (Rychlewski et al., 1999). There, each protein from the *E. coli* genome was matched against all proteins from the structural database using two sensitive sequence alignment programs, PSI-BLAST (Altschul et al., 1997) and BASIC (Rychlewski et al., 1998) and the best scoring protein was identified. As discussed in the introduction and explained in detail in Methods, these sensitive sequence comparison algorithms are used here for fold assignments, in the spirit of the threading approach. Therefore, despite using a sequence alignment algorithm, results are interpreted as fold predictions.

For every *E. coli* protein, most of the proteins from the structural database are not similar to it, and thus the alignment scores follow an extreme value distribution (Waterman, 1995). The probability that the best score is a part of this distribution (E-value) could be used as a measure of the similarity between the *E. coli* protein and the best scoring protein from the database (Altschul et al., 1997). Similar information can be conveyed by the value of the best score expressed in units of the standard deviation of the distribution (Z-score). The latter measure is better suited for the Gaussian and not extreme value distribution; nevertheless, it is often used in the literature as a measure of protein similarity. Thus, if the best score has a high Z-score (or a low E-value), the best-scoring protein is similar to the *E. coli* protein, and its structure is treated as a crude prediction of the *E. coli* protein structure. The shaded boxes in Figure 1 show the distribution of the significance of the structural prediction. In previous work (Rychlewski et al., 1999), the analysis concentrated on high significance predictions, as identified by a Z-score greater than 10 (E-value 0.05) for predictions obtained with the BASIC algorithm and an E-value less than 0.1 for predictions obtained with PSI-BLAST. BASIC and PSI-BLAST scores could not be compared directly, because they were calculated for different distributions. However, the statistical derivation of these scores is analogous, and therefore they can be compared in a qualitative way. The very conservative thresholds used for BASIC predictions were introduced to avoid the problem of false positives. This strategy allowed the identification of folds for almost 30% of the entire genome (Rychlewski et al., 1999). As discussed in the introduction, this approach had two important drawbacks. First, the fold prediction was implicitly treated as function prediction, with the underlying assumption that the function of the *E. coli* proteins should be similar to that of their putative homologous family. Several examples analyzed by hand seemed to confirm this assumption, but they represented only a small fraction of all predictions, and function similarity was not defined in any formal way. Second, with a very strict significance threshold, many correct predictions are left out of the analysis. Both problems are addressed here.

For 4,287 sequences, there are about 1,250 predictions with a Z-score greater or equal to 10 (E-value of 0.05), and there are 1,280 more with the Z-score value between 10 and 5 (E-value between 1 and 0.05). Only the former group was analyzed in the previous manuscript (Rychlewski et al., 1998); here, the entire group of predictions with Z-scores above 5 will be analyzed. Thus, the scope of the analysis almost doubles. From this group, 63% (1,593 proteins) can be matched to one of the 304 structural templates with an active site record (see Methods for details of the active site record). From this point on, only this group will be the subject of analysis. Proteins with identified active site residues form a specific subgroup of all proteins, with predominance of enzymes. The first question is how much this bias will change the score distribution. The shaded bars on Figure 2 show the number of fold predictions as a function of prediction significance for this subgroup. This part of Figure 2 is analogous to Figure 1 and suggests that the distribution of predictions in various Z-score ranges for all proteins is similar to the distribution for structures with identified functional site residues.

Each alignment from this group was evaluated using SiteMatch for conservation of residues within the active sites (see Methods for details). The numbers of proteins where more than 50% of the active site residues are conserved as a function of sequence similarity significance are shown by the solid boxes in Figure 2. As discussed in Methods, the threshold of 50% is used only for the purpose of the general analysis. In any individual case, the conservation threshold must be defined separately, often with separate rules for specific positions and specific functional signatures. Observations, such as presented here, are useful only to capture general trends. As seen in the figure, there are proteins that have conserved functional sites even if their significance scores (Z-score or E-value) are very close to a random value. For the sig-

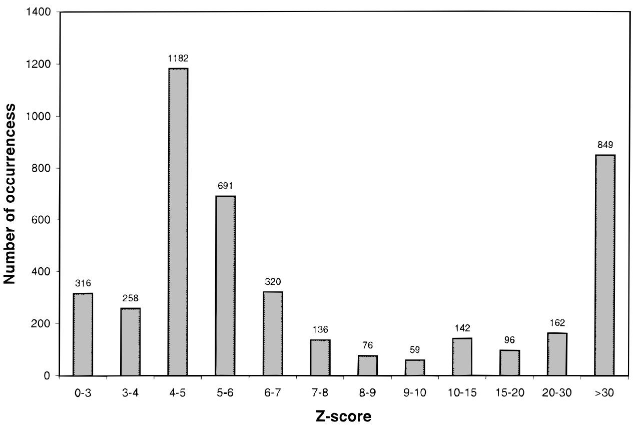


Fig. 1. The distribution of significance scores (*Z*-score) of fold predictions for proteins in *E. coli.*

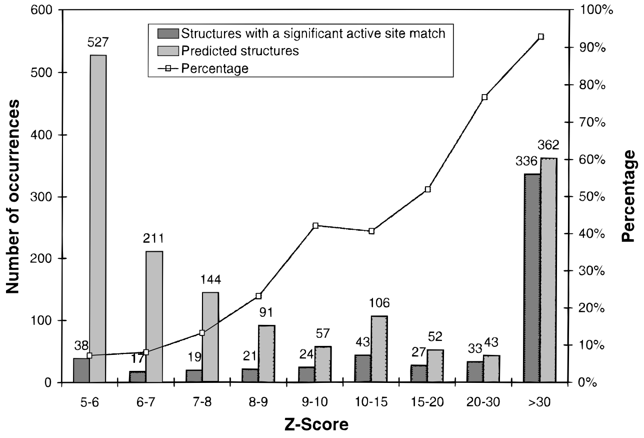


Fig. 2. SiteMatch result (active site sequence identity) vs. BASIC *Z*-scores (fold prediction) for *E. coli*. Light gray bars: total numbers of predicted folds having a SITE record in PDB file. Dark gray bars: numbers of predicted folds meeting the above criteria plus having sequence identity within SITE record greater than or equal to 50% (as defined by the BASIC alignment). Solid line: percentage of predicted folds having SITE sequence identity greater than or equal to 50%. For every *E. coli* protein, the best 5 BASIC hits exceeding a *Z*-score of 5 were taken into account. From among them, the hit with the highest residue identity in the PDB SITE record was selected (note: it was not always the best hit as defined by its *Z*-score).

nificance threshold previously used for fold predictions, about 50% of predictions have functional sites conserved. At the same time, this ratio levels off at about 90% for very similar proteins, identified with Z-scores greater than 30 (E-values less than 10^{-5}). In each of the following sections, we will concentrate separately on the two ends of the similarity spectrum. Similar trends were observed before in analysis of specific families, using a different fold assignment algorithm and different measure of prediction significance (Fetrow et al., 1998).

Predictions with significant overall similarity but a weak active site match

It is very interesting that there are proteins that do not retain their active site residues despite being closely homologous to the template. Over 124 previously made fold predictions have half or more of their active site residues missing, including 26 with prediction significance over a Z-score of 30 (see Fig. 2). In this Z-score range, the sequence identity is above 25% of identical resides, and there is no doubt that such proteins are homologous. Although it is generally accepted that the functional sites are more conserved than entire sequences, it must be pointed out that other factors, such as alignment errors, sequencing errors, and, last but not least, errors in SITE database annotations, may result in apparent nonconservation of functionally important residues.

There are four predictions with Z-scores greater than 30 where active sites are totally missed in *E. coli* proteins. (1) Three (YBCL and ORFs 01345 and 01516) are aligned with the part of the template that is not involved in its primary activity. The functional similarity between these proteins and their respective templates might be limited to other functions, not related to its primary activity as described by the SITE record. (2) The fourth (NADE) is identified as being similar to the NAD synthetase (1nsy) (the first match) or GMP synthetase (1gpmA) (the second match), and SwissProt 35 annotation identifies this protein as an NAD synthetase. The GMP synthetase has two domains. The first domain includes a conserved Cys-His-Glu and is representative of a new family of enzymes that use a catalytic triad for hydrolysis (Tesmer et al., 1996). The second domain has a nucleotide binding site that is common to the family of ATP pyrophosphatases, including NAD synthetase, asparagine synthetase, and argininosuccinate synthetase. NADE_ECOLI has only one domain that matches the second domain of GMP synthetase that does not include the active site. Interestingly, in the *E. coli* genome there are few proteins predicted to match only the first domain of GMP synthetase including two proteins with Z-scores below 10 (E-value greater than 0.05). There are no sequences that could be reliably aligned to the entire length of the known nucleotide synthetases. It is possible that in *E. coli* NAD synthetase activity is carried by an enzyme complex, rather than a single, multidomain protein.

Multidomain structures of proteins, with *E. coli* proteins being similar to nonactive domains, account for most of apparent lack of function conservation between closely homologous proteins (Z-score over 30, i.e., percent of identical residues above 25), but at lower significance levels other effects come into play. Overall, for predictions above a Z-score of 10, this effect accounts for only 20% of all cases. For proteins with Z-score values in the range between 10 and 30, sequence identity is usually below 25% of identical residues, and thus their homology is not obvious. The most important effect in this prediction significance range is due to the function diversification in structural superfamilies. For instance, the FAD/NAD-linked reductase superfamily includes glutathione and thioredoxin reductases, NADH peroxidases, and dihydrolipoamide dehydrogenases. Each of these enzymes has a different active site, and thus, a new member of this superfamily might have a function different from that of its best structural match. Over 50% of all cases of proteins with apparent lack of conservation of functional residues in the Z-score range below 30 and above 10 are predicted to belong to structural superfamilies with diverse functions. Examples with significant sequence similarity and obvious lack of conservation of active site residues account for only 20% of cases in the same significance range. An example of such a case is illustrated below, where the sequence of ORF00446 shows high sequence similarity to DNA methyltransferase (Z-score of 56.5, 25.2% sequence identity). At the same time, the active side cysteine, conserved in all known members of this family, is mutated to tryptophan.

```
IPCHRVV       lsfe
 | ||||
LPWHRVV       ORF00446
```

Thus, from proteins with high (Z-score between 30 and 10) sequence similarity and less than 50% of conserved active site residues, 50% belong to fold superfamilies with diverse functions, 20% to multidomain proteins, where only part of the protein can be matched to the *E. coli* protein and another 20% have apparently lost their activity. The remaining 10% could not be easily explained and may be a combination of alignment errors, sequencing errors, and false predictions.

Predictions with weak sequence similarity, but significant active site match

In the present prediction database, 119 predictions for the *E. coli* genome from BASIC and 29 predictions from PSI-BLAST methods have strong active site conservation, while the significance of their fold predictions is below previously used thresholds and, therefore, they were not included in the previous fold prediction list (Rychlewski et al., 1999). For these cases we can argue that conservation of active site residues could be used as additional verification of the fold prediction. Predictions in this group conform to the accepted idea that the functional residues are more conserved than protein sequences. The entire list of the new predictions for *E. coli* genome using BASIC algorithm is presented in Table 1. A few specific examples are studied below.

META_ECOLI is a homoserine transsuccinylase, predicted by the BASIC algorithm to have a similar fold to the catalytic domain of GMP synthetase. This domain includes a Cys-His-Glu triad and is representative of a new family of enzymes that use a catalytic triad for hydrolysis (Tesmer et al., 1996). The Z-score of the alignment between META and GMP synthetase is 6.82 (the sequence identity is only 17.5%), and the three active residues (C86, H181, E183) are conserved. Using PSI-BLAST, the similarity between META and GMP synthetase can also be found, but with a marginal E-value of 0.8.

ORF02791 is a hypothetical protein, predicted by BASIC to have a similar fold as a lysozyme (PDB ID: 1lzr). The Z-score is 5.49, and the sequence identity between the ORF and the lysozyme is 14.1%. The PSI-BLAST shows the significance of the ORF02791 alignment to a lysozyme is the E-value of 0.34. The residues in the

Table 1. *BASIC predictions below the significant similarity threshold (Z-score < 10) but having significant active site matches*[a]

GNO	PID	Sidn	Act	Num	Z-score	Idn	Swiss-Prot or Genebank annotation[b]	PDB annotation[b]
01648	1ajsA	11.4	SWS	1	9.450	100.0	016178 (d90811) nifs protein	Aspartate aminotransferase
00230	1amp_	13.2	SWS	1	9.630	100.0	pepd_ecoli aminoacyl-histidine dipe	Aminopeptidase (aeromonas)
00507	1amp_	12.7	SWS	1	9.280	100.0	ylbb_ecoli hypothetical 45.7 kd prot	Aminopeptidase (aeromonas)
01308	1amp_	12.9	SWS	1	5.440	100.0	(ae000231) f481; this 481 aa orf I	Aminopeptidase (aeromonas)
01309	1amp_	13.6	SWS	1	5.180	100.0	ydaj_ecoli hypothetical 47.1 kd prot	Aminopeptidase (aeromonas)
03805	1amp_	16.2	SWS	1	9.950	100.0	frvx_ecoli putative frv operon protein	Aminopeptidase (aeromonas)
04195	1amp_	13.3	SWS	1	8.910	100.0	(ae000501) hypothetical 41.7 kd prot	Aminopeptidase (aeromonas)
00774	1aqzA	16.0	CA1	4	5.050	50.0	ybia_ecoli hypothetical 18.7 kd prot	Restrictocin
00206	1bhgA	9.8	SWS	1	5.290	100.0	glo2_ecoli probable hydroxyacylglut	Beta-glucuronidase (gus gene product)
02639	1bmfD	14.3	CAT	1	5.210	100.0	reca_ecoli reca protein >gi|2098390	Bovine mitochondrial f1-atpase (f1-
00467	1broA	15.8	ACT	3	5.570	100.0	ybac_ecoli hypothetical 36.0 kd prot	Bromoperoxidase a2 (haloperoxidase)
01813	1broA	11.5	ACT	3	6.690	100.0	ptrb_ecoli protease ii (oligopeptidase)	Bromoperoxidase a2 (haloperoxidase)
02113	1broA	11.7	ACT	3	8.800	100.0	yeig_ecoli hypothetical 31.3 kd prot	Bromoperoxidase a2 (haloperoxidase)
02424	1broA	14.4	ACT	3	8.220	100.0	ypfh_ecoli hypothetical 25.7 kd prot	Bromoperoxidase a2 (haloperoxidase)
03640	1broA	14.1	ACT	3	6.100	100.0	yiel_ecoli hypothetical 44.1 kd prote	Bromoperoxidase a2 (haloperoxidase)
00084	1btl_	11.9	ACT	11	9.900	63.6	pbp3_ecoli penicillin-binding protein	Beta-lactamase tem1
00149	1btl_	8.8	ACT	11	9.900	54.5	(ae000124) penicillin-binding protein	Beta-lactamase tem1
00476	1btl_	15.4	ACT	11	6.580	54.5	ybas_ecoli hypothetical 32.9 kd prot	Beta-lactamase tem1
01495	1btl_	12.8	ACT	11	6.380	54.5	yneh_ecoli hypothetical 33.5 kd prot	Beta-lactamase tem1
02686	1ceo_	11.1	SWS	2	5.620	50.0	ygbb_ecoli hypothetical 16.9 kd prot	Cellulase celc (1,4-beta-D-glucan-
02717	1ceo_	14.7	SWS	2	5.780	50.0	ygcf_ecoli hypothetical 25.0 kd prote	Cellulase celc (1,4-beta-D-glucan-
02884	1ceo_	10.8	SWS	2	5.020	50.0	yqgf_ecoli hypothetical 15.2 kd prote	Cellulase celc (1,4-beta-D-glucan-
02927	1ceo_	11.1	SWS	2	5.660	50.0	hybd_ecoli hydrogenase-2 operon pro	Cellulase celc (1,4-beta-D-glucan-
03296	1ceo_	12.7	SWS	2	5.880	50.0	(ae000413) hypothetical 14.6 kd prot	Cellulase celc (1,4-beta-D-glucan-
02029	1csn_	11.4	SWS	1	9.120	100.0	(ae000297) f648	Casein kinase-1
03547	1csn_	13.0	SWS	1	5.480	100.0	rfay_ecoli lipopolysaccharide core	Casein kinase-1
03261	1ctn_	13.8	CA	2	5.700	100.0	yheb_ecoli hypothetical 97.1 kd prot	Chitinase a (pH 5.5, 4 °C)
02510	1ctt_	12.9	A1	4	9.960	100.0	yfhc_ecoli hypothetical 20.0 kd prot	Cytidine deaminase (cda) complexed
00347	1din_	11.5	SWS	3	8.550	66.7	yaim_ecoli hypothetical 31.4 kd prot	Dienelactone hydrolase (dlh)
01762	1fbaA	12.6	FBA	1	5.750	100.0	(ae000274) o384; uug start; this 3	Fructose-1,6-bisphosphate aldolase
01768	1fbaA	11.2	FBA	1	5.020	100.0	016319 (d90823) 3-isopropylmalate d	Fructose-1,6-bisphosphate aldolase
03788	1fbaA	18.3	FBA	1	9.860	100.0	yiht_ecoli hypothetical 32 kd protein	Fructose-1,6-bisphosphate aldolase
03084	1fds_	16.3	SWS	1	6.380	100.0	Yrar_ecoli hypothetical 24.8 kd prote	17-Beta-hydroxysteroid-dehydrogenase
03411	1fjmA	11.5	SWS	1	5.260	100.0	yhij_ecoli hypothetical 61.2 kd prote	Protein serinethreonine phosphatase-1
00299	1fxd_	3.6	FES	3	6.370	100.0	ykgf_ecoli hypothetical 53.1 kd prot	Ferredoxin ii
00870	1fxd_	9.8	FES	3	9.650	100.0	dmsb_ecoli anaerobic dimethyl sulf	Ferredoxin ii
00966	1fxd_	4.2	FES	3	7.430	100.0	yccm_ecoli hypothetical 40.1 kd pr	Ferredoxin ii
01198	1fxd_	4.1	FES	3	5.500	100.0	narh_ecoli respiratory nitrate reduc	Ferredoxin ii
01438	1fxd_	4.5	FES	3	5.910	100.0	nary_ecoli respiratory nitrate redux	Ferredoxin ii
01446	1fxd_	6.5	FES	3	7.700	100.0	fdnh_ecoli formate dehydrogenase	Ferredoxin ii
01559	1fxd_	9.8	FES	3	9.620	100.0	016034 (d90801) dimethylsulfoxid	Ferredoxin ii
01598	1fxd_	9.4	FES	3	8.980	100.0	016104 (d90806) ferredoxin ii	Ferredoxin ii
01639	1fxd_	6.3	FES	3	7.180	100.0	(u68703) hypothetical protein	Ferredoxin ii
01642	1fxd_	7.7	FES	3	7.060	100.0	(u68703) hypothetical protein	Ferredoxin ii
02162	1fxd_	5.6	FES	3	8.280	100.0	naph_ecoli ferredoxin-type protein	Ferredoxin ii
02163	1fxd_	8.7	FES	3	7.080	100.0	napg_ecoli ferredoxin-type protein	Ferredoxin ii
02166	1fxd_	10.4	FES	3	9.000	100.0	napf_ecoli ferredoxin-type protein	Ferredoxin ii
02201	1fxd_	3.8	FES	3	5.830	100.0	glpc_ecoli anaerobic glycerol-3-phos	Ferredoxin ii
02239	1fxd_	10.0	FES	3	9.080	100.0	nuoi_ecoli nadh dehydrogenase i chai	Ferredoxin ii
02432	1fxd_	8.7	FES	3	8.820	100.0	(ae000335) hypothetical 22.2 kd prot	Ferredoxin ii
02439	1fxd_	8.3	FES	3	8.380	100.0	hyfh_ecoli hydrogenase-4 component	Ferredoxin ii
02653	1fxd_	9.7	FES	3	8.430	100.0	hydn_ecoli electron transport protein	Ferredoxin ii
02660	1fxd_	12.2	FES	3	8.460	100.0	hycf_ecoli formate hydrogenlyase sub	Ferredoxin ii
02664	1fxd_	8.4	FES	3	8.650	100.0	hycb_ecoli formate hydrogenlyase sub	Ferredoxin ii
02822	1fxd_	14.7	FES	3	7.880	100.0	(u28375) orf_f163	Ferredoxin ii
02930	1fxd_	6.7	FES	3	7.170	100.0	mbht_ecoli hydrogenase-2 small chai	Ferredoxin ii
03495	1fxd_	10.8	FES	3	9.030	100.0	(ae000435) hypothetical 17.5 kd prot	Ferredoxin ii
03800	1fxd_	7.7	FES	3	7.390	100.0	fdoh_ecoli formate dehydrogenase-o,	Ferredoxin ii
03967	1fxd_	6.7	FES	3	7.980	100.0	nrfc_ecoli nrfc protein >gi|2144352	Ferredoxin ii
04047	1fxd_	8.6	FES	3	5.490	100.0	frdb_ecoli fumarate reductase iron-	Ferredoxin ii
04266	1fxd_	5.9	FES	3	5.880	100.0	yjjw_ecoli hypothetical 31.5 kd prot	Ferredoxin ii

(*continued*)

Table 1. *Continued.*

GNO	PID	Sidn	Act	Num	Z-score	Idn	Swiss-Prot or Genebank annotation[b]	PDB annotation[b]
02720	1gpmA	13.5	GAA	3	9.600	100.0	pyrg_ecoli ctp synthase	gmp synthetase (xmp aminase)
03908	1gpmA	13.4	GAA	3	6.820	100.0	(u00006) homoserine transsuccinylas	gmp synthetase (xmp aminase)
00714	1iba_	5.0	S1	1	9.100	100.0	hrsa_ecoli hrsa protein >gi|2121156	Glucose permease fragment
02840	1iyu_	18.6	LIP	1	7.070	100.0	gcsh_ecoli glycine cleavage system	Dihydrolipoamide acetyltransferase
00511	1jdc_	13.4	SWS	3	5.070	66.7	ylbf_ecoli hypothetical 29.6 kd prot	1,4-Alpha maltotetrahydrolase
00742	1jud_	15.1	CAT	2	5.140	100.0	ybha_ecoli hypothetical 30.2 kd prot	l-2-Haloacid dehalogenase
01919	1jud_	16.8	CAT	2	6.820	100.0	(ae000287) o271; this 271 aa orf I	l-2-Haloacid dehalogenase
03128	1jud_	13.1	CAT	2	9.210	50.0	(ae000399) hypothetical 20.0 kd prot	l-2-Haloacid dehalogenase
00901	1lbu_	16.2	CAT	3	8.600	100.0	ycbk_ecoli hypothetical 20.4 kd pro	Muramoyl-pentapeptide
01366	1ldg_	13.0	SWS	1	8.240	100.0	015717 (d90777) 3-hydroxybutyryl-co	l-Lactate dehydrogenase
02299	1ldg_	12.9	SWS	1	7.350	100.0	(ae000322) f714; this 714 aa orf I	l-Lactate dehydrogenase
03759	1ldg_	11.2	SWS	1	8.340	100.0	fadb_ecoli fatty oxidation complex	l-Lactate dehydrogenase
02236	1lml_	11.4	ACT	5	6.330	60.0	nuol_ecoli nadh dehydrogenase i cha	Leishmanolysin (gp63 protein, psp)
00813	1lt3A	12.4	SWS	1	5.090	100.0	(ae000186) o371; this 371 aa orf I	Heat-labile enterotoxin fragment (lt-i)
02791	1lzr_	20.3	SWS	2	5.490	100.0	(u28375) orf_o138 escherichia coli	Lysozyme (lz406) complexed with
03662	1nhp_	14.3	SWS	2	6.280	50.0	(ae000451) glucose inhibited divis	Nadh peroxidase (npx) mutant with
02086	1pamA	8.7	SWS	3	5.160	66.7	yehv_ecoli hypothetical transcripti	Cyclodextrin glucanotransferase
03578	1pamA	12.5	SWS	3	9.670	66.7	(ae000443) hypothetical 88.1 kd pr	Cyclodextrin glucanotransferase
04087	1pii_	12.6	ASS	5	8.580	60.0	sgah_ecoli probable hexulose-6-phos	*N*-(5′phosphoribosyl)anthranilate
02990	1pth_	14.1	PER	3	6.680	66.7	cca_ecoli trna nucleotidyltransfera	Prostaglandin h2 synthase-1
03020	1scuA	11.0	SWS	1	8.270	100.0	(ae000390) o334; sequence change j	Succinyl-coa synthetase (succinate-coa
03785	1smd_	14.6	SWS	3	6.040	66.7	(ae000463) hypothetical 77.2 kd pr	Amylase
02024	1svpA	13.4	TRI	3	6.800	66.7	dcd_ecoli deoxycytidine triphosphat	Sindbis virus capsid protein fragment
01747	1thtA	10.8	CAT	3	7.230	66.7	g3p1_ecoli glyceraldehyde 3-phosph	Thioesterase
00825	1thx_	11.9	DIS	2	7.440	100.0	glr1_ecoli glutaredoxin 1 (grx1) >g	Thioredoxin (thioredoxin 2)
01624	1thx_	19.0	DIS	2	6.090	50.0	ydhd_ecoli hypothetical 12.9 kd prot	Thioredoxin (thioredoxin 2)
04012	1tplA	14.0	PLA	7	7.930	57.1	6 arginine decarboxylase, biodegrad	Tyrosine phenol-lyase
00410	1vid_	11.6	SWS	2	7.000	50.0	pgpa_ecoli phosphatidylglycerophosp	Catechol o-methyltransferase (comt)
02939	1vid_	11.7	SWS	2	5.940	50.0	exbd_ecoli biopolymer transport exb	Catechol o-methyltransferase (comt)
03212	1vid_	13.8	SWS	2	8.070	50.0	sun_ecoli sun protein (fmu protein)	Catechol o-methyltransferase (comt)
00110	1vjs_	10.9	SWS	3	5.550	66.7	ampd_ecoli ampd protein >gi|78310|p	Alpha-amylase (bla)
00379	1vjs_	9.6	SWS	3	5.140	66.7	(ae000145) o192; 100 pct identical	Alpha-amylase (bla)
01120	1vjs_	9.4	SWS	3	8.040	66.7	(ae000214) o189; phage stats; this	Alpha-amylase (bla)
01330	1vjs_	12.7	SWS	3	6.340	66.7	(ae000233) o285; this 285 aa orf I	Alpha-amylase (bla)
01781	1vjs_	12.4	SWS	3	5.380	66.7	yeab_ecoli hypothetical 21.4 kd prot	Alpha-amylase (bla)
02259	1vjs_	11.1	SWS	3	6.180	66.7	yfcf_ecoli hypothetical 24.3 kd prote	Alpha-amylase (bla)
02275	1vjs_	12.6	SWS	3	7.210	66.7	deda_ecoli deda protein (dsg-1 proein	Alpha-amylase (bla)
02721	1vjs_	14.1	SWS	3	5.170	66.7	mazg_ecoli mazg protein >gi|882675	Alpha-amylase (bla)
02881	1vjs_	10.9	SWS	3	5.040	66.7	(u28377) orf_o252	Alpha-amylase (bla)
03894	1vjs_	10.2	SWS	3	5.910	66.7	yjae_ecoli hypothetical 18.2 kd prote	Alpha-amylase (bla)
00941	1vsd_	12.3	ACT	3	5.800	66.7	(ae000198) f122; this 122 aa orf I	Integrase fragment
01741	1xyzA	11.8	SWS	2	5.160	100.0	016287 (d90820) tagatose-bisphosp	1,4-Beta-ᴅ-xylan-xylanohydrolase
00585	2af8_	11.9	S42	1	5.880	100.0	entb_ecoli isochorismatase (2,3 dih	Actinorhodin polyketide synthase acyl
02093	2bltA	11.4	CTA	7	9.510	57.1	pbp7_ecoli penicillin-binding prote	Beta-lactamase (cephalosporinase)
01437	2btfA	12.6	CAT	3	5.860	66.7	narw_ecoli respiratory nitrate redu	Beta-actin-profilin complex
02028	2btfA	13.1	CAT	3	9.190	66.7	(ae000297) o471; uug start; 99 pct	Beta-actin-profilin complex
01592	2dkb_	12.4	PLP	1	9.760	100.0	maly_ecoli maly protein >gi|96164|p	2,2-Dialkylglycine decarboxylase
01980	2dkb_	15.5	PLP	1	9.740	100.0	his8_ecoli histidinol-phosphate amin	2,2-Dialkylglycine decarboxylase
02839	2dkb_	12.1	PLP	1	9.050	100.0	gcsp_ecoli glycine dehydrogenase	2,2-Dialkylglycine decarboxylase
03494	2dkb_	16.1	PLP	1	8.670	100.0	avta_ecoli valine–pyruvate aminotr	2,2-Dialkylglycine decarboxylase
04026	2dkb_	13.7	PLP	1	7.020	100.0	dcly_ecoli lysine decarboxylase	2,2-Dialkylglycine decarboxylase
00161	2sga_	10.3	SWS	3	7.670	100.0	heat shock protein	Proteinase a (component of the
03164	2sga_	10.5	SWS	3	9.670	100.0	degq_ecoli protease degq precursor	Proteinase a (component of the
00766	3dni_	11.3	ACT	4	6.280	100.0	(ae000181) f253	Deoxyribonuclease i (dnase i)
00164	4kbpA	11.3	SWS	1	7.790	100.0	(u70214) hypothetical protein	Purple acid phosphatase
02210	4pgmA	17.0	CIC	4	8.120	75.0	ais_ecoli ais protein >gi|1788586	Phosphoglycerate mutase 1
03165	5ptp_	14.6	CAT	4	8.180	100.0	degs_ecoli protease degs precursor	Beta trypsin
00015	5rubA	13.9	SWS	1	5.460	100.0	dnaj_ecoli dnaj protein >gi|72228	Rubisco (ribulose-1,5-bisphosphate)

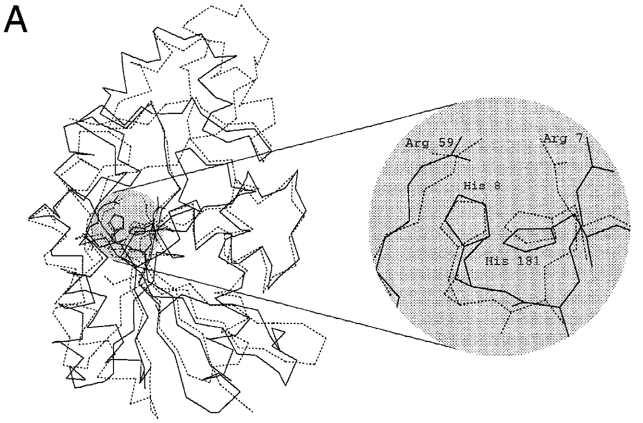
[a] GNO, gene number; PID, the identification of a structure; Sidn, sequence identity between a sequence and a structure; Num, number of active site residues; Act, names of actives sites, except "SWS" for information coming from Swiss-Prot, others use names from "SITE" records in PDB; Idn, sequence identity in active site.

[b] Descriptions in the two last columns of Table 1 were shortened to fit into the manuscript format. The full text of Table 1 is available at the WEB site bioinformatics.burnham-inst.edu.

active site (E35, D53) are conserved in both the BASIC and PSI-BLAST alignments.

AIS_ECOLI was predicted to have a similar fold to that of 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (PDB ID: 1bif) and phosphoglycerate mutase 1 biological unit (PDB ID: 4pgmA). The Z-score of the alignments are 8.66 and 8.12, while the sequence identity between AIS and the two proteins with known structures are 12.0 and 17.0%, respectively. The first structure does not have a SITE record. Checking the alignment between the ORF and the second structure by SiteMatch, three out of four active site residues are conserved. The fourth one is a histidine that is aligned with a gap. Investigating the structural alignment between the first and the second fold predictions, it can be seen that the histidine in the active site from both structures is actually aligned as shown in Figure 3. Meanwhile, the alignment between AIS and the first fold prediction shows that a histidine from AIS is aligned with the same histidine as in the structural alignment. Thus, it can be concluded that the alignment between AIS and 4gpmA contains errors, and in fact, all residues in the active site should be aligned. This example illustrates that by combining SiteMatch and information from structural alignments, one can find and correct errors in specific alignments.

Five proteins (YBAC, PTRB, YEIG, YPFH, and YIEL) are predicted to have a similar fold and exact active site match as bromoperoxidase a2 (PDB ID: 1broA), with a Z-score below 10.0 (an additional eight *E. coli* proteins are predicted to have the same fold with a Z-score above 10). Bromoperoxidase catalyzes the bromination of organic compounds in the presence of bromide and peroxide. The overall structure of bromoperoxidase can be characterized as an α/β-hydrolase fold with a catalytic triad of Ser532, D617, and H652 (Hecht, 1994). The sequence identities between all five *E. coli* proteins, and 1broA are listed in Table 2. Four of them are hypothetical proteins. Only one is known as protease II (PTRB), which catalyzes the hydrolysis of Arg-|-XAA and Lys-|-XAA bonds in oligopeptides, even when P1′ is proline (Kanatani et al., 1991). Although the significance score (Z-score) between PTRB and 1broA is only 6.9, it has the same active triad (Ser532, D617, and H652) as the template. The reactive serine residue of protease II was experimentally identified as Ser532 (Kanatani et al., 1991). The sequence around the serine residue is identical to the common sequence of Gly-X-Ser-X-Gly, which has been found in the active site of most serine proteases, thus function prediction based on local patterns might misclassify it as serine protease. Except for this region, protease II showed no significant sequence similarity with *E. coli* serine protease, protease IV, and protease La (Kanatani et al., 1991), and consequently, it is usually classified as a member of a separate family, the prolyl oligopeptidase family of peptidases. A recent study suggests a striking secondary structure similarity between serine carboxypeptidase and prolyl oligopeptidase (Medrano et al., 1998). This observation indirectly supports our prediction in this study.



B

```
? 1bif_ 119.1.1.1.2.1 6-phosphofructo-2-kinase fructose-2,6-bisphosphatase Mutant
biological_unit
LIGANDS:
* CAT  220  289  354
1 ATG   10   11   12   13   14   15   16   17   94  120  131  134  135  136
       184  210  391
2  MG   15   16   92
3 PO4  219  220  226  269  289  354  355
4 PO4  300  314  318  329  355  359
5 GOL  263  287  339  342  343
6  S1    9   10   11   12   13   14   15   16
7  S2   88   89   90   91   92
========
ALIGNMENT:
        66 6666661
CPTLIVMVGL PARGKTYISK KLTRYLNFIG VPTREFNVGQ YRRDMVKTYK SFEFFLPDNE EGLKIRKQCA
---------- ---------- ---------- ---------- ---------- ---------- ----------
                  777 77 1                           1            1  111
LAALNDVRKF LSEEGGHVAV FDATNTTRER RAMIFNFGEQ NGYKTFFVES ICVDPEVIAA NIVQVKLGSP
---------- ---------- ---------- ---------- ---------- ---------- ----------

                                                         1
DYVNRDSDEA TEDFMRRIEC YENSYESLDE EQD------- --RDLSYIKI MDVGQSYVVN RVADHIQSRI
---------- ---------- ---------- ---MLAFCRS SLKSKKYIII LLALAAIAGL GTHAAWSSNG
        1                3*      3                                        5
VYYLMNIHVT P------RSI YLCRHGESEL NLKGR-IGGD PGLSPRGREF SKHLAQFISD QNIKDLKVFT
LPRIDNKTLA RLAQQHP-VV VLFRHAERCD RSTNQCLSDK TGITVKGTQD ARELGNAF-S ADIPDFDLYS
    3                   5 *           4              4    4           4
SQMKRTIQTA EALSVPYEQF KVLNEIDAGV CEEMTYEEIQ DHYPLEFALR DQDKYRYRYP KGESYEDLVQ
SNTVRTIQSA TWFS------ ---------- --AGKKLTV- ---DKRLLQC GNEIYS---- ----------
    5  55               *4   4                                     1
RLEPVIMEL- --ERQENVLV ICHQAVMRCL LAYFLD--KA AEELPYLKCP LHTVLKLTPV AYGCKVESIF
----AIKDLQ SKAPDKNIVI FTHNHCLTYI AKDKRDATFK PDYLDGLVM- ---------- --HVEKGKVY

LNVAAVN-TH RDRPQNVDIS RPSEEALVTV PAHQ
LDGEFVNH-- ---------- ---------- ----
==========
NOTE  :        align   pdb w_idt a_idt __ligand_  zscore ratio ident simil
RESULT:        02210 1bif_  12.0  13.1 * CAT   3    8.66  0.62  66.7  66.7
RESULT:        02210 1bif_  12.0  13.1 1 ATG  17    8.66 -0.35   0.0   5.9
RESULT:        02210 1bif_  12.0  13.1 2  MG   3    8.66 -0.50   0.0   0.0
RESULT:        02210 1bif_  12.0  13.1 3 PO4   7    8.66  0.68  57.1  71.4
RESULT:        02210 1bif_  12.0  13.1 4 PO4   6    8.66 -0.29   0.0  16.7
RESULT:        02210 1bif_  12.0  13.1 5 GOL   5    8.66  0.48  20.0  60.0
RESULT:        02210 1bif_  12.0  13.1 6  S1   8    8.66 -0.42   0.0   0.0
RESULT:        02210 1bif_  12.0  13.1 7  S2   5    8.66 -0.43   0.0   0.0
-----------------------------------------------------------------------------

? 4pgmA 119.1.1.1.2.1 phosphoglycerate mutase 1 biological_unit
LIGANDS:
* CIC    8  181    7   59
========
ALIGNMENT:
                                                                * *
---------- ---------- ---------- ---------- ---------- ---PKLVLVR HGQSEWNEKN
MLAFCRSSLK SKKYIIILLA LAAIAGLGTH AAWSSNGLPR IDNKTLARLA QQHPVVVLFR HAERCDRSTN
                                             *
LFTGWVDVKL SAKGQQEAAR AGELLKEKKV YPDVLYTSKL SRAIQTANIA LEKADRLWIP VNRSWRLNER
QCLS-DKTGI TVKGTQDARE LGNAF-SADI PDFDLYSSNT VRTIQSA-TW FSAGKKLTVD KRLLQCGNEI

HYGDLQGKDK AETLKK---- ---------- ---------- ---------- ---------- ----FGEEKF
-YSAIKDLQS KAPDKNIVIF THNHCLTYIA KDKRDATFKP DYLDGLVMHV EKGKVYLDGE FVNH------

NTYRRSFDVP PPPIDASSPF SQKGDERYKY VDPNVLPETE SLALVIDRLL PYWQDVIAKD LLSGKTVMIA
---------- ---------- ---------- ---------- ---------- ---------- ----------
 *
AHGNSLRGLV KHLEGISDAD IAKLNIPTGI PLVFELDENL KPSKPSYYLD PEAAAA
---------- ---------- ---------- ---------- ---------- ------
==========
NOTE  :        align   pdb w_idt a_idt __ligand_  zscore ratio ident simil
RESULT:        02210 4pgmA  17.0  22.3 * CIC   4    8.12  0.67  75.0  75.0
```

Fig. 3. A: Superposition of the C_α trace of 6-phosphofructo-2-kinase fructose-2,6-bisphosphatase (1bif) and phosphoglycerate mutase 1 (4pgm, chain A). 1bif is shown as a dashed line and 4pgmA as a solid line. The inset is the superposition of the enlarged active sites. **B:** Annotated alignments for fold predictions of gene 02210. The abbreviations used here are the same as those in Figure 7. The histidine discussed in the paper is shown in bold.

Conservation of the hydrophobic pattern

One of the most important and, at the same time, most difficult to study experimentally, aspects of protein function are their interactions with other proteins. Assembly of multienzyme complexes, interactions between regulatory proteins and their targets, all depend critically on the character of the interacting surfaces. For many proteins, the function prediction could not be complete without prediction of the way they associate with other proteins. At the same time, the mode of association often varies in homologous families, often correlating with the changes in function.

Our experience in studying the domain assembly within the family of calcium binding EF-hand proteins (Pawlowski et al., 1996) showed that analysis of conservation and position of surface residues can be successfully used in predicting a mode of multi-

Table 2. *Sequence identities between 5 ORFs and 1broA*

Sequence identity	PTRB	YEIG	YPFH	YIEL	1broA
YBAC	15.5	18.8	15.8	18.7	15.8
PTRB		12.4	14.6	16.7	11.5
YEIG			16.9	15.9	11.7
YPFH				17.3	14.4
YIEL					14.1

meric assembly of multidomain proteins. Proteins from this family are built from 70 amino acid domains, which could be found as monomers (ICaBPs used for calcium storage), dimers [S100B proteins, participating (among other functions) in signal processing in the brain] or domains in multidomain proteins (calmodulin, troponin C, recoverin, and several other proteins). Here we present an application of these ideas to automated analysis of protein interfaces.

In particular, the conservation or changes in patches of hydrophobic residues on the surface could be used as an indication of the similarity in the mode of multimeric assembly of the predicted protein to that of its best-scoring structural template. All fold predictions were analyzed for hydrophobic pattern conservation, as described in Methods. **S**, the sum of absolute values of the burial energy difference between the template sequence in its own structure and the prediction target sequence in the template structure, was calculated for all fold predictions. The distribution of **S** is shown in Figure 4. As expected, the distribution of energy differences is very broad, suggesting a very strong level of noise from random mutations. The sudden widening of the distribution around Z-score 5 (*E*-value of 1), where most of the fold predictions and alignments become essentially random, is clearly visible. At the same time, several proteins clearly lay outside of the envelope of the distribution. The complete list of such proteins is available from the authors' Web server. Here, a few examples will be analyzed in more detail.

First examples nicely illustrate the predictive possibilities of burial energy difference analysis. The spermidine binding protein strongly recognizes (Z-score of 56.5) the structure of the maltodextrin binding protein, despite a low sequence identity of 16% (Fig. 5). At the same time, the hydrophobic energy difference for the target protein in the template structure is 61.4 energy units, which on the per-residue basis is almost three times higher than the average. The template, maltodextrin binding protein is a monomer in solution, while the experimental structure for a spermidine binding protein (known, but absent from our database) is known to be an octamer. Another example is the cysB transcription factor. It is weakly recognized to be similar to the lysine binding protein, with a Z-score of 6.9 and the sequence identity of 17.2%, but an unusually high hydrophobic energy difference of 78.6 energy units. This prediction was independently validated by experiments when the cysB protein from *Klebsiella aerogens* was crystallized. This structure, now available in the PDB, was absent from our database of structural templates, which was prepared before the cysB structure submission. The cysB protein has an unusual tetrameric structure in contrast to the lysine binding protein, which is a monomer.

The next example attempts to predict the oligomerization state in addition to fold and function prediction. The hypothetical protein YBCK_ECOLI is predicted to be homologous to a diphtheria toxin. The alignment has a Z-score of 37, indicating a very strong match despite a low sequence similarity of 12%. At the same time, the first 70 residues, containing the DNA binding site and two out of five catalytic residues, are missing from the sequence of the *E. coli* protein. Two other catalytic residues are conserved, with a third one probably missed due to an alignment error (see Fig. 7). The hydrophobic pattern from a dimeric diphtheria toxin, present

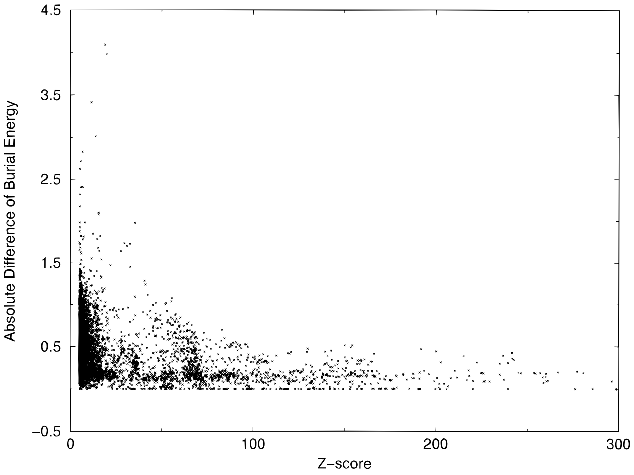


Fig. 4. Absolute difference of burial energy per aligned residue vs. significance scores (*Z*-score) for all predictions with a *Z*-score greater than 5 in *E. coli*. Energy units correspond to kT in room temperature (Godzik et al., 1992).

```
################################   01096  ################################
? 1pot_ 272.1.1.1.3.2 spermidinePUTRESCINE-BINDING PROTEIN (potd)
LIGANDS:
1 SPD    9    10    11    12    58    60   143   146   204   230   232   268   302
========
ALIGNMENT:
                                    1111
---------- ---------- -----NNTLY FYNWTEYVPP GLLEQFTKET GIKVIYSTYE SNETMYAKLK
MKKWSRHLLA AGALALGMSA AHADDNNTLY FYNWTEYVPP GLLEQFTKET GIKVIYSTYE SNETMYAKLK
             1 1
TYKDGAYDLV VPSTYYVDKM RKEGMIQKID KSKLTNFSNL DPDMLNKPFD PNNDYSIPYI WGATAIGVNG
TYKDGAYDLV VPSTYYVDKM RKEGMIQKID KSKLTNFSNL DPDMLNKPFD PNNDYSIPYI WGATAIGVNG
                             1   1
DAVDPKSVTS WADLWKPEYK GSLLLTDDAR EVFQMALRKL GYSGNTTDPK EIEAAYNELK KLMPNVAAFN
DAVDPKSVTS WADLWKPEYK GSLLLTDDAR EVFQMALRKL GYSGNTTDPK EIEAAYNELK KLMPNVAAFN
                   1                            1 1
SDNPANPYME GEVNLGMIWN GSAFVARQAG TPIDVVWPKE GGIFWMDSLA IPANAKNKEG ALKLINFLLR
SDNPANPYME GEVNLGMIWN GSAFVARQAG TPIDVVWPKE GGIFWMDSLA IPANAKNKEG ALKLINFLLR
             1                                    1
PDVAKQVAET IGYPTPNLAA RKLLSPEVAN DKTLYPDAET IKNGEWQNDV GAASSIYEEY YQKLKAG-
PDVAKQVAET IGYPTPNLAA RKLLSPEVAN DKTLYPDAET IKNGEWQNDV GAASSIYEEY YQKLKAGR
==========
NOTE      :     zscore    ENG0    LEN0    ENG1    LEN1  ALNLEN  ABSDIF
TOTENERGY:      124.94  -48.95     322  -48.95     348     322    0.00
EXPENERGY:      124.94  -23.87     322  -23.87     348     322    0.00
INFENERGY:      124.94   -7.12     322   -7.12     348     322    0.00
BURENERGY:      124.94  -17.96     322  -17.96     348     322    0.00
NOTE  :         align   pdb w_idt a_idt __ligand_  zscore ratio ident simil
RESULT:         01096 1pot_  92.5 100.0 1 SPD  13  124.94  1.00 100.0 100.0
---------------------------------------------------------------------------

? 1omp_ 272.1.1.1.3.1 D-maltodextrin-binding protein
LIGANDS:
========
ALIGNMENT:

KIE------- ---------- ------EGKL VIWINGDKGY NGLAEVGKKF EKDTGIKVTV EHPDKLEEKF
---MKKWSRH LLAAGALALG MSAAHADDNN TLYFYNWTEY -VPPGLLEQF TKETGIKVIY STYESNETMY

PQVAATGDGP -DIIFWAHDR FGGYAQSGLL AEITPDKAFQ DKLYPFTWDA VRYN-GKLIA YPIAVEALSL
AKLKTYKDGA YDLVVPSTYY VDKMRKEGMI QKIDKSKLTN FSNLDPDMLN KPFDPNNDYS IPYIWGATAI

IYNKDLL-PN PPKTWEEIPA LDKELKAKGK SALMFNLQEP YFTWPLIAAD GGYAFKYENG KYDIKDVGVD
GVNGDAVDPK SVTSWADLW- ---KPEYKGS LLLTDDAREV FQMALRKLGY SGNT------ --------TD

NAGAKAGLTF LVDLIKN-KH MNADTDYSIA EAAFNKGETA MTINGPWAWS NIDTSKVNYG VTVLPTFKGQ
PKEIEAAYNE LKKLMPNVAA FNSDNP---- ANPYMEGEVN LGMI--WNGS AFVARQAGTP IDVVWPKEG-

PSKPFVGVLS AGINAASPNK ELAKEFLENY LLTDEGLEAV NKDK---PLG AVALKSYEEE LAKDPRIAAT
---GIFWMDS LAIPANAKNK EGALKLIN-F LLRPDVAKQV AETIGYPTPN LAARKLLSPE VANDKTLYPD

MENAQKGEIM PNIPQMSAFW YAVRTAVINA -ASGRQTVDE ALKDAQTRIT K
AETIKNGEWQ NDVGAASSIY EEYYQKLKAG R--------- ---------- -
==========
NOTE      :     zscore    ENG0    LEN0    ENG1    LEN1  ALNLEN  ABSDIF
TOTENERGY:       56.49  -61.39     370  -16.48     348     317  108.25
EXPENERGY:       56.49  -21.53     370  -13.07     348     317   19.94
INFENERGY:       56.49   -8.69     370   -4.60     348     317   21.97
BURENERGY:       56.49  -31.17     370    1.19     348     317   66.34
```

Fig. 5. Fold predictions for gene 01096 in *E. coli*. The abbreviations used here are the same as those in Figure 7.

in our template database, gives a very high burial energy, suggesting that the YBCK protein has a different oligomerization state.

Discussion

In fold prediction, the goal is to assign a fold to a protein sequence by finding a most compatible fold from a library of known protein folds. Here, this goal was achieved by using a supersensitive sequence alignment program, which for every protein from the *E. coli* genome identified the most similar protein sequence from a group of proteins with known structures. In previous work, as well as in examples shown in Methods, it was verified that such similarity, if sufficiently strong as measured by the statistical significance of the similarity score, indeed translates into the fold similarity (Rychlewski et al., 1999). This way, a new protein can be (putatively) included in the fold superfamily. But how to extend the fold prediction to function prediction? The most common approach is to assume that the function of a putative superfamily member is going to be the same as the already known members of the family. However, this type of "implicit" prediction may not work when the evolutionary distance between the new protein and the known proteins increases to the point where function divergence becomes common. It becomes even more questionable, when the homology relation becomes uncertain. One easy way to confirm or refute this type of naive prediction is to check the conservation of the active site or ligand binding site residues. Traditionally, this approach concentrates on specific protein families. With thousands of fold predictions available on genome scale (Casari et al., 1996; Fischer & Eisenberg, 1997; Jones, 1998; Pawlowski et al., 1999; Rychlewski et al., 1998, 1999), the automated alignment analysis becomes increasingly important.

An automated method to verify the conservation of the functional site residues for alignments from sequence analysis and fold prediction methods was used to analyze the results of the previous fold prediction for proteins from the *E. coli* genome. Using SiteMatch, thousands of detailed function prediction verifications can be done in a few hours. The results presented here offer several insights into the common practice of using distant sequence similarity, which can be verified as fold prediction, for function prediction. Even for clearly homologous proteins with significant sequence identity, about 10% do not contain conserved functional site residues. This percentage drops to 50% for weakly similar proteins, where the relationships between proteins are variously interpreted as distant homology or accidental fold similarity. Although a part of this effect can be explained by alignment errors, sequencing errors, database annotation errors, or other trivial explanations, clearly the effect of function diversification remains and is likely to challenge many existing functional assignments in genomic databases. Detailed analyses of specific protein families provided examples of this problem (Fetrow et al., 1998). Here, the large-scale analysis of over 300 protein families gives a large-scale view that confirms insights obtained from smaller scale studies.

At the other end of the significance spectrum, the functional site conservation analysis offers a simpler and more immediate application. Conservation of functional features could strengthen many fold predictions with low significance scores. For prediction in this significance range, the ratio between the distant homologues and accidental fold similarities shifts toward the latter and the number of incorrect fold assignments becomes significant. Over a hundred new predictions with low sequence similarity but strong conservation of functional features may be added to the prediction list, raising the number of fold/homology predictions to over 32% of all *E. coli* proteins. We can expect this number to increase even more as fold prediction methods are improved, and more function descriptions become available. This marks a qualitative increase over the usually quoted 10–15% of the genome proteins for which folds could be predicted (Casari et al., 1996; Fischer & Eisenberg, 1997; Frishman & Mewes, 1997). It strongly suggests that for proteins from newly sequenced genomes fold predictions, followed by detailed function predictions will play an increasing role in interpretation of the genomic information.

Analysis of the hydrophobic pattern conservation in homologous proteins, while very preliminary, also illustrates additional insights into the biological function of proteins identified in genomic studies that could be achieved with fold assignments. Several examples discussed in this paper, together with over 100 additional ones available from the authors' Web server, identify proteins that change the way they associate in complexes, which opens a way to more detailed analysis of their function.

The key to the success of such automated analysis is the quality of the database of the detailed functional information about the known protein structures. Due to the incomplete annotation in PDB, it is impossible to assign functional information for every protein structure. Also, the SITE record is not used very consistently. For instance, SITE records often include binding sites, which often change with changing specificity of binding, even if the activity is constant. Such information is useful to obtain a general view of function conservation in homologous families, but for detailed study of specific families, functional information from detailed structural analysis of specific structures must be used instead of database derived information. Such function signatures for several protein functions were built and used in detailed predictions for specific functions (Fetrow et al., 1998; Fetrow & Skolnick, 1998).

Methods

Protein structural database and group of representative structures

The vast majority of experimentally known protein structures is deposited in the Brookhaven PDB (Bernstein et al., 1977). The release from January 1998, including 6,700 proteins that have over 11,000 single chains, has been used in the work described here. Many of the proteins were solved multiple times, and for many others, structures of their close homologues were solved and deposited in the PDB as well. Thus, to avoid multiple counting of protein structures, all PDB single chains were clustered based on their sequence similarity. The goal is to divide all proteins within PDB into groups, such that all proteins within one group would have a similarity above the threshold, and all pairs with each of the proteins coming from a different cluster have similarity below threshold. Several such sets were prepared by different groups, with the most popular sets being available from EMBL (Hobohm et al., 1992). In this work, we use a set prepared at EMBL, based on a 30% sequence identity threshold. Each structure in the set is accompanied with a cluster of similar proteins, which can be used to crosscheck and compare function-related annotations. There are 1,151 proteins in the set; the complete list is available from the authors' Web site as pdb30. According to our definition, different function would mean completely different active sites.

An analysis of a similar set of sequence clusters (results not shown) showed that there are no clusters with completely different active sites. All the clusters but four had at least 50% sequence conservancy in the active site, and majority of clusters had 100% conservancy.

Database of multilevel function signatures

As mentioned in the introduction, the term "function" is used in many different meanings, often encompassing such different concepts as activity, mechanism of action, or function in the organization of entire organisms. As a first approximation to such a multilevel function description, we have decided to focus on three aspects of function: (1) activity, as described by active site residues in an enzyme; (2) ligand binding, as described by the residues in binding sites and residues in contact with inhibitors, cofactors, etc. cocrystallized with the protein; and (3) interaction with other proteins, as described by a pattern of hydrophobic residues on the surface.

Such description is obviously highly simplified and does not attempt to provide a complete functional description of proteins in the structural database. The SITE database contains information from three sources: (1) SITE records of PDB files: most of the protein structural files are annotated by their authors with annotations identified by specific keywords following PDB guidelines. In particular, the "SITE" record is intended to describe residues involved in biological activity. This information was extracted directly from the PDB files and reformatted into a specific format used later by the SiteMatch program. About 500 of the 1,151 proteins in the structural database (see the previous section) have at least one protein with a "SITE" record in their homologous cluster. At the same time, structure depositors have significant freedom in including various residues into the SITE record. To arrive at a more consistent definition, the residues in the SITE record of each PDB file are cross-checked with additional information coming from the PDB file or other databases: E.C. classification, presence of specific keywords in protein name or MEDLINE record, residue conservation in the immediate homologous family, and others. (2) Functional annotations in the SwissProt 34 sequence database: the curated protein sequence database contains information about active site residues. This information was extracted using a simple script. (3) Analysis of ligand (prosthetic groups, substrates, or inhibitors) binding in PDB structures: such groups are denoted as HETATOM in the PDB files, and residues involved in their binding can be identified by searching for all protein atoms that are within a certain cutoff distance from any of the HETATOMs in a PDB file. Here, a 3.9 Å cutoff is used for all ligand atoms, including DNA and RNA. Names of ligands are extracted from the PDB files and include all HETATOM records except water molecules.

When combining information from all sources, some functional annotation can be made for about 705 proteins out of 1,151 in our structural database. For 304 proteins the SITE database identifies active site residues. This latter group was the focus of the analysis presented here.

Finally, surface regions involved in interactions between different proteins can be identified by the presence of hydrophobic residues on the protein surface. To identify such sites, information about the burial/exposed status of all positions along the sequence and corresponding statistical potential parameters were adopted from the topology fingerprint description of protein structures (Godzik et al., 1992; Jaroszewski et al., 1998).

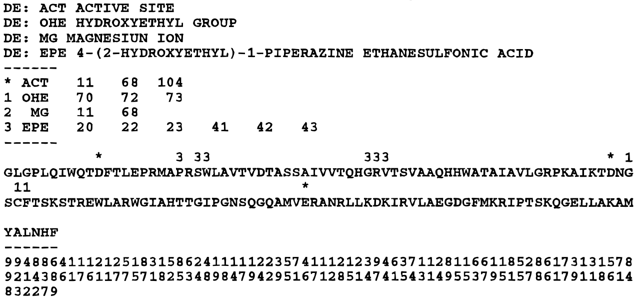
An example of a record in the SITE database is presented in Figure 6. A star denotes the active sites extracted from "SITE" records and confirmed as described above. Consecutive numbers beginning with "1" denote the ligand binding sites. In most proteins, multiple ligand binding sites are identified. A list of identified functionally important positions is presented, following the name of the ligand or a star symbol for the active site. Finally, a list of buried/exposed/interface assignments for all positions along the sequence is included in the SITE database record. The SITE database can be extended, and its records can be improved and verified by literature searching, cross-reference with other similar databases (Laskowski et al., 1997) or detailed analysis of functional requirements in the experimental structure (Fetrow & Skolnick, 1998). Various improvement of the database are planned for the near future, and the fold and function prediction database for several microbial genomes, available at the authors' Web site at bioinformatics.burnham-inst.org uses a continuously updated super set of the features described here.

Fold assignments

Fold assignments for proteins from *E. coli* genome were adapted from a preceding manuscript (Rychlewski et al., 1999). In this and other related manuscripts (Pawlowski et al., 1999; Rychlewski et al., 1998, 1999), two algorithms were used for fold assignments. The first one is a profile-to-profile comparison method BASIC (Rychlewski et al., 1998). The second is the position specific iterated BLAST (PSI-BLAST) algorithm (Altschul et al., 1997), which is the newest version of the de facto standard of database protein similarity searching algorithms.

Both algorithms are able to detect weak sequence similarities, in many cases not detectable using standard methods of sequence analysis. It is not clear what is the exact relation between proteins with such similarities. As shown in the next section, it can be shown that for proteins with sequence similarity above certain threshold, folds of both proteins (if known) are always similar. Thus, both algorithms are used as fold predictors, similar in spirit to application of threading algorithms.

For both algorithms, a representative subset of all proteins in the PDB, as described in the first section of Methods, is used as a database of potential templates. This is the same set that was used



```
>1vsd_ E.C.2.7.7.49 Integrase Fragment: catalytic core domain
DE: ACT ACTIVE SITE
DE: OHE HYDROXYETHYL GROUP
DE: MG MAGNESIUN ION
DE: EPE 4-(2-HYDROXYETHYL)-1-PIPERAZINE ETHANESULFONIC ACID
------
* ACT   11   68  104
1 OHE   70   72   73
2  MG   11   68
3 EPE   20   22   23   41   42   43
------
          *        3 33                 333                        * 1
GLGPLQIWQTDFTLEPRMAPRSWLAVTVDTASSAIVVTQHGRVTSVAAQHHWATAIAVLGRPKAIKTDNG
 11                              *
SCFTSKSTREWLARWGIAHTTGIPGNSQGQAMVERANRLLKDKIRVLAEGDGFMKRIPTSKQGELLAKAM

YALNHF
------
9948864111212518315862411111223574111212394637112811661185286173131578
9214386176117757182534898479429516712851474154314955379515786179118614
832279
```

Fig. 6. A sample record from the SITE database for avian sarcoma virus integrase (PDB ID: 1vsd). In the lines marked with "DE": the active site, the site of a hydroxyethyl group (bound covalently to a cysteine), a magnesium binding site and HEPES binding site are listed. These sites are denoted by "ACT," "OHE," "MG," and "EPE," respectively. The hydrophobic pattern follows the reference sequence of the protein.

in the previous work (Rychlewski et al., 1999). The best-scoring protein from the database is identified for each of the proteins from the *E. coli* genome. The alignment between these two proteins is used as an input to the SiteMatch program. Because of the specific choice of the database (proteins with known structures), BASIC and PSI-BLAST programs that employ only sequence information are used as fold predictors.

Alignments for all target–template pairs with a significance score above a certain threshold are used. Fold predictions for two genomes (*M. genitalium* and *E. coli*) with all methods used in this manuscript are available on the authors' Web server at bioinformatics.burnham-inst.org. A third genome, *H. pylori*, is analyzed with updated versions of the template database and upgraded fold prediction algorithms (Pawlowski et al., 1999) and is now also available on the server.

Optimization and verification of the BASIC algorithm

The BASIC algorithm was optimized to recognize the maximal number of structurally similar proteins on benchmarks customized for fold prediction algorithms. A particular benchmark available from the Web server at UCLA (http://www.doe-mbi.ucla.edu/people/frsvr) was used during the development of a BASIC algorithm. This benchmark consists of 68 target proteins for which the correct template (structural similar protein) has to be found in a database of about 300 examples. The results (Table 3) presented here show that a sequence-only fold recognition method can closely match the prediction accuracy of best threading algorithms. A more extensive evaluation of different fold recognition algorithms is presented elsewhere (L. Rychlewski, L. Jaroszewski, K. Pawlowski, A. Godzik, in prep.).

Site identification

SiteMatch is a computer program designed to analyze the conservation of residues in functionally important regions in target–template alignments. It uses the SITE database described above and the alignment between the new protein (the prediction target) and an already characterized protein (the template). The alignments from various sources can be used, including, but not limited to, the BLAST, BASIC, and threading methods.

The criterion for significant site match, used later in the statistical analysis of function conservation for the entire genome, is 50% of sequence identity in functional site. This threshold was introduced rather arbitrarily, only for the purpose of general analysis. In every specific case, the threshold must be chosen individually, and often different positions must be treated in a different way. If there is no active site information available for the first hit of a prediction, the second hit will be checked, and so on, until a template with a functional description is found.

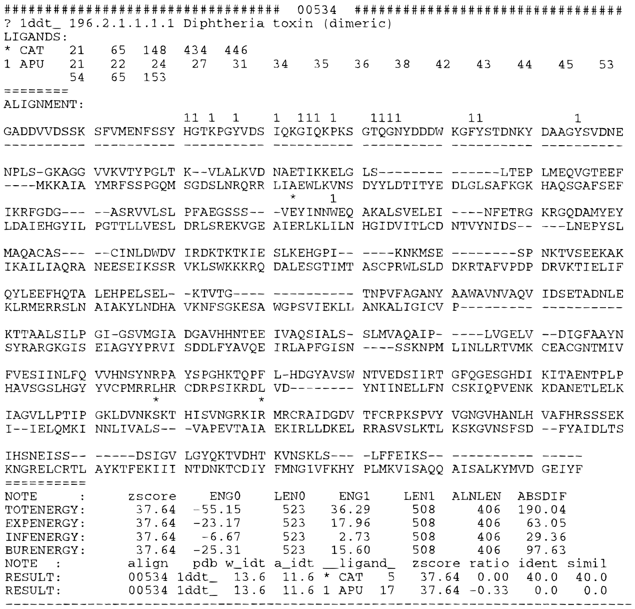
Because the functional description in the SITE database usually involves the active site and/or one or more binding sites, conservation can be calculated separately for each separate record. This allows a more complete functional conservation analysis, because often only some of several functional records are conserved in the alignment.

The pattern of buried/exposed positions along the sequence is used to assess the conservation of interprotein interactions, which is important in multimeric assembly. Interaction sites between proteins often could be recognized as patches of hydrophobic residues on protein surfaces. Often, even closely related proteins assemble in different complexes, which is reflected in the different positions

Table 3. *Results achieved on the UCLA threading benchmark containing 68 target-template pairs and a database of 300 templates*[a]

	Rank = 1	Rank ≤ 5	Rank ≤ 10
Simple BLAST	27	—	—
PSI-BLAST	32	—	—
THREADING			
(Godzik et al., 1992)	22	30	34
Global sequence alignment	40	50	52
THREADING			
(Jaroszewski et al., 1998)	54	58	60
BASIC	52	57	60

[a] The values present the number of pairs, where the template obtained a rank given above. For BLAST predictions it is difficult to estimate lower significance predictions, because they often are not listed due to a large number of homologous proteins.



```
##################################  00534  ##################################
? 1ddt_ 196.2.1.1.1.1 Diphtheria toxin (dimeric)
LIGANDS:
* CAT    21    65   148   434   446
1 APU    21    22    24    27    31    34    35    36    38    42    43    44    45    53
         54    65   153
========
ALIGNMENT:
                      11 1  1    1  111 1    1111        11           1
GADDVVDSSK SFVMENFSSY HGTKPGYVDS IQKGIQKPKS GTQGNYDDDW KGFYSTDNKY DAAGYSVDNE
---------- ---------- ---------- ---------- ---------- ---------- ----------

NPLS-GKAGG VVKVTYPGLT K--VLALKVD NAETIKKELG LS-------- ------LTEP LMEQVGTEEF
----MKKAIA YMRFSSPGQM SGDSLNRQRR LIAEWLKVNS DYYLDTITYE DLGLSAFKGK HAQSGAFSEF
                                   *    1
IKRFGDG--- --ASRVVLSL PFAEGSSS-- -VEYINNWEQ AKALSVELEI ----NFETRG KRGQDAMYEY
LDAIEHGYIL PGTTLLVESL DRLSREKVGE AIERLKLILN HGIDVITLCD NTVYNIDS-- ---LNEPYSL

MAQACAS--- --CINLDWDV IRDKTKTKIE SLKEHGPI-- ----KNKMSE --------SP NKTVSEEKAK
IKAILIAQRA NEESEIKSSR VKLSWKKKRQ DALESGTIMT ASCPRWLSLD DKRTAFVPDP DRVKTIELIF

QYLEEFHQTA LEHPELSEL- -KTVTG---- ---------- TNPVFAGANY AAWAVNVAQV IDSETADNLE
KLRMERRSLN AIAKYLNDHA VKNFSGKESA WGPSVIEKLL ANKALIGICV P--------- ----------

KTTAALSILP GI-GSVMGIA DGAVHHNTEE IVAQSIALS- SLMVAQAIP- ----LVGELV --DIGFAAYN
SYRARGKGIS EIAGYYPRVI SDDLFYAVQE IRLAPFGISN ----SSKNPM LINLLRTVMK CEACGNTMIV

FVESIINLFQ VVHNSYNRPA YSPGHKTQPF L-HDGYAVSW NTVEDSIIRT GFQGESGHDI KITAENTPLP
HAVSGSLHGY YVCPMRRLHR CDRPSIKRDL VD-------- YNIINELLFN CSKIQPVENK KDANETLELK
                  *            *
IAGVLLPTIP GKLDVNKSKT HISVNGRKIR MRCRAIDGDV TFCRPKSPVY VGNGVHANLH VAFHRSSSEK
I--IELQMKI NNLIVALS-- -VAPEVTAIA EKIRLLDKEL RRASVSLKTL KSKGVNSFSD --FYAIDLTS

IHSNEISS-- -----DSIGV LGYQKTVDHT KVNSKLS--- -LFFEIKS-- ---------- -----
KNGRELCRTL AYKTFEKIII NTDNKTCDIY FMNGIVFKHY PLMKVISAQQ AISALKYMVD GEIYF
==========
NOTE      :     zscore    ENG0    LEN0    ENG1    LEN1  ALNLEN  ABSDIF
TOTENERGY:       37.64  -55.15     523   36.29     508     406  190.04
EXPENERGY:       37.64  -23.17     523   17.96     508     406   63.05
INFENERGY:       37.64   -6.67     523    2.73     508     406   29.36
BURENERGY:       37.64  -25.31     523   15.60     508     406   97.63
NOTE  :         align   pdb w_idt a_idt __ligand_  zscore ratio ident simil
RESULT:         00534 1ddt_  13.6  11.6 * CAT   5   37.64  0.00  40.0  40.0
RESULT:         00534 1ddt_  13.6  11.6 1 APU  17   37.64 -0.33   0.0   0.0
------------------------------------------------------------------------------
```

Fig. 7. The output from the SiteMatch program for assessing the alignment between gene 00534 and template (1ddt). The alignment is annotated according to the functional information of the template. The one body energy (denoted by "TOTENERGY") is divided into three terms, exposed (EXPENERGY), interface (INFENERGY), and burial (BURENERGY). "ENG0" is calculated by putting the sequence of the template into its own structure while "ENG1" by putting the query sequence into the template according to the alignment between them. "LEN0" is the length of the template, and "LEN1" is the length of the query sequence. "ALNLEN" is the length of the aligned residues. "ABSDIF" is the absolute energy difference summed up in residue level. In the "RESULT" part, "w_idt" is sequence identity between the query sequence and the template while "a_idt" only in aligned region. Active site and ligand information is given under "__ligand_"; the similarity "ratio" is calculated by dividing the alignment score in a site by the self-alignment score of residues in the "ident" and "simil" are the sequence identity and similarity in functional sites, respectively.

of such hydrophobic patches (Pawlowski et al., 1996). Here, such differences are estimated by calculating **S,** the sum of absolute values of the difference between the burial energy of the prediction target sequence and the original native sequence, using the template burial/exposed pattern according to the alignment. The summation is over all residues in the alignment. Any significant changes in the number and the position of hydrophobic residues on positions exposed to the solvent would result in the substantial changes to the value of **S**.

The result of a SiteMatch analysis of the target–template alignment is quite complex, listing conservation of various functional sites and changes in burial/exposed energy. An example of such analysis is presented in Figure 7, with several other examples presented earlier in this paper. The main purpose of such analysis is to provide a starting point in a detailed evaluation of the function conservation for every fold prediction. Records similar to that presented in Figure 7 are a part of a genomic prediction web site maintained in our group. For general analysis, such as presented in this paper, only some features of a full analysis were used.

Database availability

The fold prediction database for the proteins from the *M. genitalium, E. coli*, and *H. pylori* genomes is available on the group's Web server at http://bioinformatics.burnham-inst.org (also available at http://cape6.scripps.edu) as described previously (Rychlewski et al., 1999), is now enhanced with function predictions.

Acknowledgments

This work was supported by NSF Grant MCB-9506278 (B.Z., L.R., K.P., and A.G.), and NIH Grant No. GM48835 (J.S. and J.F.).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acid Res 25*:3389–3402.

Bairoch A, Apweiler R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl Acid Res 27*:49–54.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Simanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bork P, Gibson TJ. 1996. Applying motif and profile searches. *Methods Enzymol 266*:162–184.

Bowie JU, Luethy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three dimensional structure. *Science 253*:164–170.

Casari G, Ouzonis C, Valencia A, Sandr C. 1996. GeneQuiz: Automatic function assignment for genome sequence analysis. In: *Proceedings of the first annual Pacific symposium on biocomputing*. Hawaii: World Scientific. pp 108–119.

Fetrow J, Godzik A, Skolnick J. 1998. Functional analysis of the *E. coli* genome using the sequence-to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J Mol Biol 282*:703–711.

Fetrow J, Skolnick J. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxin/thioredoxin and T1 ribonuclease. *J Mol Biol 281*:949–968.

Fischer D, Eisenberg D. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc Natl Acad Sci USA 94*:11929–11934.

Frishman D, Mewes H. 1997. Protein structural classes in five complete genomes. *Nat Struct Biol 4*:626–628.

Godzik A, Skolnick J, Kolinski A. 1992. A topology fingerprint approach to the inverse folding problem. *J Mol Biol 227*:227–238.

Gribskov M, McLachlan M, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355–4358.

Hecht HJ. 1994. The metal-ion-free oxireductase from *Streptomyces aureofaciens* has an α/β hydrolase fold. *Struct Biol 1*:532–537.

Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci 1*:409–417.

Huynen M, Dandekar T, Bork P. 1998. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett 426*:1–5.

Jaroszewski L, Rychlewski L, Zhang B, Godzik A. 1998. Fold prediction by a hierarchy of sequence and threading methods. *Protein Sci 7*:1431–1440.

Jones D. 1998. GenTHREADER. Warwick, http://globin.bio.warwick.ac.uk/genome.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.

Kanatani A, Matsuda T, Shimoda T, Misoka F, Lin X, Yoshimoto T, Tsuru D. 1991. Protease II from *Escherichia coli*: Sequencing and expression of the enzyme gene and characterization of the expressed enzyme. *J Biochem 110*:315–320.

Laskowski R, Hutchinson G, Michie A, Wallace A, Jones M, Martin A, Luscombe N, Milburn D, Thornton J. 1997. PDBsum: A WEB-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci 22*:488–490.

Medrano F, Alonso J, Garcia J, Romero A, Bode W, Gomis-Ruth F. 1998. Structure of protein iminopeptidase from *Xanthomonas campestris* pv. citri: A prototype for the prolyl oligopeptidase family. *EMBO J 17*:1–9.

Murzin AG. 1998. How far divergent evolution goes in proteins. *Curr Opin Struct Biol 8*:380–387.

Pawlowski K, Bierzynski A, Godzik A. 1996. Structural diversity in a family of homologous proteins. *J Mol Biol 258*:349–366.

Pawlowski K, Rychlewski L, Zhang B, Godzik A. 1999. The *Helicobacter pylori* genome: From sequence analysis to structural and functional predictions. *Proteins*. In press.

Pearson WR, Miller W. 1992. Dynamic programming algorithms for biological sequence comparison. *Methods Enzymol 210*:575–601.

Russell RB, Copley RR, Barton GJ. 1996. Protein fold recognition by mapping predicted secondary structures. *J Mol Biol 259*:349–365.

Rychlewski L, Zhang B, Godzik A. 1998. Function and fold predictions for *Mycoplasma genitalium* proteins. *Folding Design 3*:229–238.

Rychlewski L, Zhang B, Godzik A. 1999. Functional insights from structural predictions: Analysis of *Escherichia coli* genome. *Protein Sci 8*:614–624.

Tesmer J, Klem T, Deras M, Davisson V, Smith J. 1996. The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two new enzymatic families. *Nat Struct Biol 3*:74–86.

Waterman MS. 1995. *Introduction to computational biology: Maps, sequences and genomes (interdisciplinary statistics)*. London: Chapman & Hall.