Protein Science (1999), *8*:614–624. Cambridge University Press. Printed in the USA.

Functional insights from structural predictions: Analysis of the *Escherichia coli* genome

LESZEK RYCHLEWSKI, BAOHONG ZHANG, AND ADAM GODZIK
Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

(RECEIVED July 21, 1998; ACCEPTED November 19, 1998)

Abstract

Fold assignments for proteins from the *Escherichia coli* genome are carried out using BASIC, a profile–profile alignment algorithm, recently tested on fold recognition benchmarks and on the *Mycoplasma genitalium* genome and PSI BLAST, the newest generation of the de facto standard in homology search algorithms. The fold assignments are followed by automated modeling and the resulting three-dimensional models are analyzed for possible function prediction.

Close to 30% of the proteins encoded in the *E. coli* genome can be recognized as homologous to a protein family with known structure. Most of these homologies (23% of the entire genome) can be recognized both by PSI BLAST and BASIC algorithms, but the latter recognizes an additional 260 homologies. Previous estimates suggested that only 10–15% of *E. coli* proteins can be characterized this way. This dramatic increase in the number of recognized homologies between *E. coli* proteins and structurally characterized protein families is partly due to the rapid increase of the database of known protein structures, but mostly it is due to the significant improvement in prediction algorithms.

Knowing protein structure adds a new dimension to our understanding of its function and the predictions presented here can be used to predict function for uncharacterized proteins. Several examples, analyzed in more detail in this paper, include the DPS protein protecting DNA from oxidative damage (predicted to be homologous to ferritin with iron ion acting as a reducing agent) and the ahpC/tsa family of proteins, which provides resistance to various oxidating agents (predicted to be homologous to glutathione peroxidase).

Keywords: genome analysis; protein structure predictions

The most important tool in predicting structures and functions of newly determined proteins is built on a simple observation that homologous proteins have similar folds and strong similarities in their functions. Therefore, establishing homology to an already known and characterized protein group is usually the first step in the analysis of a new protein. Once the homology is established, it is possible to make various inferences about the structure, activity, and function of the new protein.

For closely related proteins, the homology recognition is simple because sequences retain a significant level of similarity. The structures of such proteins remain similar and the structure of one protein can be used as a template to build a model of the second. Most importantly, the function of close homologues rarely changes and that means that function prediction is easy.

The situation is much more complicated for distantly related proteins. Here, even the existence of the evolutionary relationship is often disputable. This problem started receiving more attention with the discovery of many protein groups with similar folds, but without any apparent sequence similarity (Orengo et al., 1993). Recognition of such proteins before structure determination could be viewed as a "limited protein structure prediction." It could not substitute for a general solution to a folding problem, but it has a very significant practical application. Thus, a new class of structure prediction methods, termed "inverse folding" or "threading," has been specifically formulated to find such structural similarities.

In an inverse folding approach, one "threads" the sequence of a new, uncharacterized protein (prediction target) through different template structures and attempts to find the most compatible structure. In the last few years, many algorithms have been developed (Bowie et al., 1991; Godzik et al., 1992; Jones et al., 1992; Bryant & Lawrence, 1993; Ouzounis et al., 1993; Matsuo & Nishikawa, 1994; Yi & Lander, 1994; Wilmanns & Eisenberg, 1995; Alexandrov et al., 1996; Russell et al., 1996; Jaroszewski et al., 1998). In our laboratory, we have developed a fold prediction hierarchy where several sequence/sequence, sequence/structure, structure/structure alignment algorithms and modeling tools are combined to create a fully automated prediction protocol (Jaroszewski et al., 1998). This protocol starts with a sequence and, subject to a successful prediction step, produces a full three-dimensional model of the prediction target (Jaroszewski et al., 1998; Rychlewski et al., 1998). The model is subsequently analyzed, both for independent predic-

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

tion verification, and for possible function prediction. Different algorithms specifically search for different classes of structural similarity—distant homology or random structural similarity. By enhancing sequence information with position dependent similarity scoring obtained from the analysis of multiple alignments of closely related sequences, combined with information about the structure or predicted structure, our prediction hierarchy achieves a much higher recognition rate than standard sequence homology analysis tools (Jaroszewski et al., 1998; Rychlewski et al., 1998). Despite all the differences, the classical paradigm of homology modeling is followed with its three basic steps of identifying the structural template, creating the alignment, and building the model. Because of large evolutionary distances between targets and templates, it is impossible to strive for high, atomic level accuracy in structure prediction. Instead, the models are analyzed for plausibility (whether or not they can be built at all) and for conservation of features important for function. This is the first step in the direction of detailed functional prediction.

All algorithms from the fold prediction hierarchy were tested on fold recognition benchmarks, where fold prediction of proteins with no detectable sequence similarity is attempted, hoping to find structurally similar protein in the database (see Methods). In fold recognition benchmark, such predictions are done afterward, when the correct answer is known. Surprisingly, the super-sensitive sequence alignment algorithms, such as PSI BLAST (Altschul et al., 1997) or BASIC (Rychlewski et al., 1998), can almost match the fold recognition ratio of threading algorithms using structural data (see Methods), while being faster and easier to use. Therefore, both methods were applied to do the fold assignments for proteins from a genome of the simple pathogenic bacteria *Mycoplasma genitalium* (Rychlewski et al., 1998). About 38% of the proteins coded by the *M. genitalium* genome could be assigned to a protein family with an already characterized structure. This represents an over twofold increase over previously published fold assignments for this genome (Casari et al., 1996; Fischer & Eisenberg, 1997; Frishman & Mewes, 1997), and about a 40% improvement over one iteration of the position specific iterative BLAST (Altschul et al., 1997), the most recent generation of the widely used sequence analysis program. Fifty new (as compared to PSI BLAST) structure predictions were made, suggesting a structural framework for several proteins with known functions and predicting several new biochemical mechanisms and activities in *M. genitalium* (Rychlewski et al., 1998). In this paper, we continue this work by applying PSI BLAST and BASIC algorithms to the proteins from the *Escherichia coli* genome.

E. coli is a Gram-negative bacterium that inhabits the lower gut of animals, including humans. It is a favorite model organism for biochemical, molecular biology, and genetic studies. This is mostly because it is easy to work with, has simple nutritional needs, and displays interesting behavior, including conjugation and growth of many different viruses. It can grow on a minimal medium, and unlike many of the organisms whose genomes were sequenced earlier, it has a fairly complete set of metabolic pathways. Interest in this organism was further enhanced by the emergence of several virulent strains that were responsible for several well-publicized outbreaks. For these various reasons, it is the most studied organism on earth and provides most of the insights we have about the organization of life (Moxon & Higgins, 1997). Elucidation of its entire genome was one of the most anticipated events in modern biology. Although it was not the first genome to be completed, its analysis is a blueprint for analyses of other genomes and organisms.

The first part of Results describes statistics of fold predictions using different methods. The second part of Results focuses on function verification based on the analysis of alignments between prediction targets and the templates identified in the first step of the analysis, as well as on the analysis of the three-dimensional models built on those alignments by an automatic modeling procedure. The third part of Results follows a few examples to a deeper functional analysis level. The rest of the predictions are available at cape6.scripps.edu. All techniques and algorithms used in this manuscript are described in Methods.

Results

The set of 4,287 protein sequences from the *E. coli* genome was downloaded from the *E. coli* genome web site at the University of Wisconsin, Madison (www.genetics.wisc.edu). Each of these sequences was compared to a superset of all public protein sequence databases (see Methods) using the PSI BLAST algorithm (Altschul et al., 1997). Output from PSI BLAST is used to prepare the input for subsequent stages of analysis, but it is also a powerful prediction tool on its own. In the sequence database, all proteins with known three-dimensional structures are identified with the Protein Data Base (PDB) keyword. Therefore, a keyword search in the PSI BLAST output reveals all PSI BLAST-recognized homologies to the proteins with known structures. In the next step, the *E. coli* sequences were compared to a smaller database containing a set of proteins representing all currently known protein folds. The BASIC program from the suite of fold prediction algorithms developed in our laboratory was used (Jaroszewski et al., 1998; Rychlewski et al., 1998) to identify the best templates for all *E. coli* proteins. The alignments for all target-template pairs for which the score was better than the predefined threshold were used for automated modeling. The MODELLER program (Sali, 1994) was used in this step of the analysis. Statistics of the fold assignments are discussed in the following paragraph and presented in Table 1. Technical details about the algorithms, databases, and protocols for fold assignments are discussed in Methods. The entire database of structural predictions and the resulting models is available at cape6.scripps.edu. The remainder of the paper is devoted to the analysis of these data. In particular, we concentrate on the analysis of the most reliable predictions, those with E-values for predictions with the BASIC algorithm below 0.05. The E-value (see Methods) is a number of random similarities with the given score, expected by chance. In tests on fold recognition benchmarks, no false predictions were found with the E-values above this threshold.

Table 1. *The number of high significance structural predictions for proteins from the E. coli genome*[a]

Prediction method	Number of high significance predictions
BLAST version 1.3.2	497
PSI BLAST version 2.0.2	990
BASIC version 1.1	1,250

[a] Significance of 0.1 E-value was used as a threshold for both versions of the BLAST algorithm, and 0.05 E-value was used for the BASIC algorithm.

Predictions with lower significance are often correct; therefore, additional predictions and models are also presented on cape6.scripps.edu, but assignment of their accuracy is difficult. For instance, the similarity between the lysR family of transcriptional regulators and periplasmic binding proteins is recognized with the E-value of 0.5. This similarity was confirmed by a recently solved structure of cysB from *Klebsiella aerogens* (Tyrrell et al., 1997), which was not incorporated into our structural database at the time the calculations described here were made. The structure, predicted by the procedure described in this paper, is compared in Figure 1 to the now available crystal structure. This case offers an interesting example of function diversification in families of homologous proteins. At the same time, it gives a natural "reality check" of what can be predicted in the case of a successful prediction.

The periplasmic binding proteins (PBP) can be found in the space between the cytoplasmic (i.e., inner) membrane and the cell

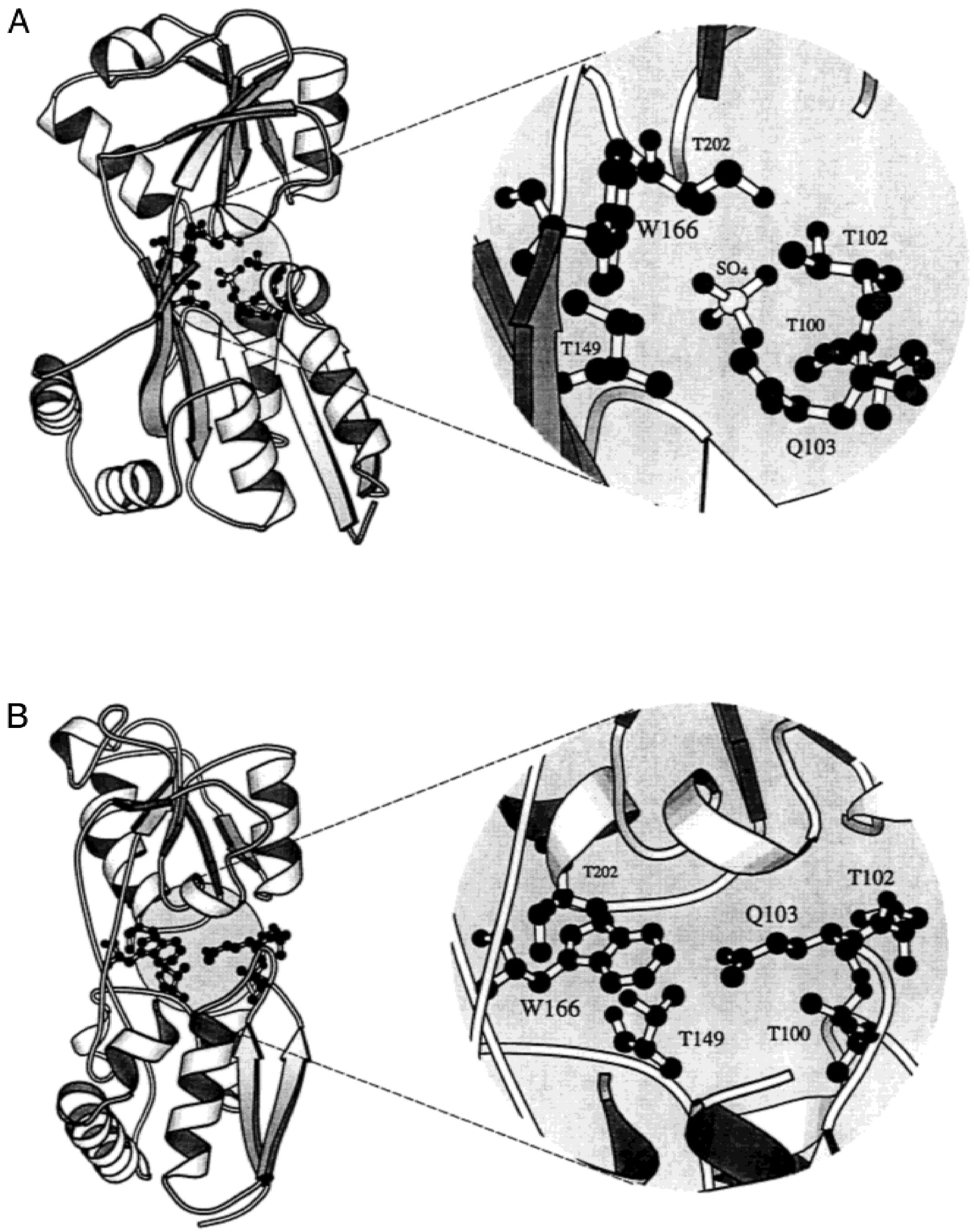


Fig. 1. The comparison of (**A**) an experimental structure of cofactor binding fragment of cysB protein from *Klebsiella aerogenes* (PDB structure code 1al3) and (**B**) the model of its *E. coli* homologue, prepared by an automated recognition/modeling procedure described in this paper. For both the structure and the model, the enlarged active (binding) site is shown in an inset. The sequence similarity between cysB from *K. aerogenes* and *E. coli* is 92% identical residues.

wall of Gram-negative bacteria such as *E. coli*. PBPs are the initial receptors of the active transport systems for carbohydrates, amino acids, and ions, or receptors in a chemotactic response (Quiocho & Ledvina, 1996). CysB belongs to a large family of cofactor-binding repressor/operon domains, such as *lacR* and *amiC*. In Figure 1, our automated prediction procedure resulted in a model with many realistic features. Despite a relatively large difference from the crystal structure (6.5 Å root-mean-square deviation (RMSD) for all Cα atoms in the model), this model correctly identified all residues involved in substrate binding (inset, Fig. 1). This is not a small achievement because the binding site is the most variable part of the sequence and the overall sequence similarity between the closest PDB and cysR is 17% sequence identity. At the same time, the binding site is severely distorted with an RMSD of residues involved in binding between a model and a crystal structure exceeding 3 Å. An error of this magnitude would make docking simulations or detailed substrate/inhibitor design impractical. On the other hand, the general features of a binding site, such as its global charge and type of residues involved in binding, are predicted correctly. Parenthetically, there is also an enzyme, porphobilinogen deaminase, that has the same topology (SCOP, 1995) and whose relation to the PBP family is recognized by the BASIC algorithm.

For the 4,279 protein sequences, the PSI BLAST algorithm (see Methods) detected 990 significant (E-value lower than 0.1) similarities to proteins with known structures. This constitutes 23% of the entire genome, a ratio close to but lower than the 27% obtained with the identical procedure for the *M. genitalium* genome (Rychlewski et al., 1998). Similar decreases are seen for predictions with other prediction methods (see below). The decrease is probably due to the "minimal" character of the *M. genitalium* genome, containing only the most fundamental and better-studied protein functions.

The BASIC program from our fold-recognition suite detected 1,250 significant (E-value lower than 0.05) similarities to proteins with known structures, an over 25% increase over the PSI BLAST recognition rate. Again, the percentage of recognized proteins (29%) is smaller than in the case of *M. genitalium* proteins where this percentage was equal to 38%. Predictions with the BASIC algorithm form a superset of BLAST predictions because all but four above-threshold significance made by PSI BLAST are recognized by the BASIC algorithm. All four cases can be traced to independent recognition of separate domains in a multidomain protein. The significance threshold of the E-value of 0.05 used for BASIC predictions (see Methods) is rather conservative. For instance, there are an additional 300 predictions, none of them recognized by PSI BLAST above its significance threshold, with E-values lower than 0.1. On fold recognition benchmarks, where the BASIC algorithm was tested, there were no false predictions with E-values lower than 1.6, so there is a good chance that many predictions with a lower significance level are actually accurate.

From the 1,250 structural predictions made by the BASIC program, 190 (124 of which are also recognized by PSI BLAST) represent hypothetical proteins with unknown functions. Therefore, for this group, the structural predictions represent a first step toward function predictions.

Presentation of over 200 predictions is difficult and even a brief discussion of all new predictions would greatly extend this paper. Tables 1 and 2 present a sample of predictions for proteins with known functions and for hypothetical proteins, respectively. Only the group of predictions not recognized by the PSI BLAST algorithm is represented in both tables, because in most cases they represent novel and more interesting predictions.

Analysis of structural predictions

Verification of fold predictions, such as presented in Tables 1 and 2, is difficult because the structures are not known. These are genuine predictions that are done with an automated prediction server. However, the fold predictions done here are in fact based on an assumption of the homology between the prediction target and the structural template. Therefore, analysis of functional similarities could be used to further validate the possible homology. Such analysis depends on how much is known about the function of the prediction target.

Specific function of the prediction target is known

For proteins whose function is known, usually from experiment, the analysis of the models can provide verification of the structural prediction. A small sample of predictions of this type is given in Table 1, and examples are discussed below. Comparing the functions of *E. coli* proteins to those of the structural superfamilies identified in the prediction can make a first level of verification. Analysis of the model, specifically checking the conservation of

Table 2. *A sample from over 250 high significance predictions for proteins from the E. coli genome*[a]

E-value	PDB code of a template		Target	
0.05	1ofgA	Oxidoreductase	MHPF	Acetaldehyde dehydrogenase
0.03	1ukz_	Uridylate kinase	GNTV	Thermosensitive gluconokinase
0.02	1bcfA	Bacterioferritin	DPS	DNA protection protein
0.02	4xis_	Xylose isomerase	SGAU	Hexulose
2e−4	1gp1A	Peroxidase	BCP	BCP protein
2e−4	1xvaA	Methyltransferase	FTSJ	Cell division protein
2e−4	1ukz_	Uridylate kinase	AROK	Shikimate kinase
1e−4	2tys	Tryptophan synthase	RPE	Ribulose-phosphate 3-epimerase
1e−4	1xvaA	Methyltransferase	GIDB	Cell division protein
1e−4	1znbA	Lactamase	PHNP	PHNP protein

[a] SwissProt codes and functional assignments for the *E. coli* proteins are given. None of the targets in this table had a PSI BLAST hit to a protein with known structure with the E-value less than 10.0.

residues responsible for biochemical activity, provides clues to the function of the new protein. For instance, conservation of the active site residues and changes in residues involved in substrate binding is a good indication that the model is correct, especially if both proteins are known to have the same function but different specificity. Three examples below illustrate these situations:

1. Acetaldehyde dehydrogenase (MHPF) is predicted to have a fold similar to that of glucose/fructose oxidoreductase (1ofg); the same structural family also contains *E. coli* glyceraldehyde dehydrogenase. In addition to several known homologous proteins, two additional proteins from the *E. coli* genome, hypothetical protein YJHC and USG-1 protein, are predicted to have the same fold. It is interesting to note that the *E. coli* enzyme (MHPF) has a (predicted) different fold from eukariotic aldehyde dehydrogenase (reductase) despite a very similar activity.
2. Gluconokinase and shikimate kinases are predicted to have the same fold as uridilate kinase. The similarity in the ATP binding domain is much larger than in the catalytic domain. This prediction is also made as a low significance PSI BLAST prediction.
3. Hexulose-6-phosphate isomerase is predicted to be similar to xylose isomerase.

In all these cases, there is significant analogy between prediction targets and proposed template activities to support possible homology between proteins in each pair. In each case, the active site residues are conserved in the alignment, and their position on the model supports the known activity of the target protein.

General, but not specific, function is known

Predictions for proteins whose function, but not activity, is known, are particularly interesting. The examples below were chosen randomly from the list of almost 300 novel predictions.

1. The DPS protein, which protects DNA from oxidative damage during prolonged starvation, is predicted to be similar to bacterioferritin. This prediction was recently confirmed by experiment where a protein from *Listeria innucua*, which is homologous to the *E. coli* DPS protein, was shown to behave as ferritin by being able to oxidize and sequester iron ions (Bozzi et al., 1997).
2. The PHNP protein, thought to be involved in alkylphosphonate uptake and degradation, is predicted to have the fold of metallo-β-lactamase.
3. One of the proteins from the *fts* gene, the ftsJ protein involved in cell division, is predicted to be similar to glycine n-methyltransferase.
4. A protein involved in antioxidant resistance and homologous to a subunit of hydroperoxidate reductase (bacterioferritin comigratory protein (BCP)) is predicted to be similar to glutathione peroxidase. This prediction could be extended to an entire AHPC/TSA family whose members provide resistance to various oxidating agents, such as sulfur radicals, hydroperoxide, and thiolperoxide.

In all these examples, structural prediction allows one to make a prediction of specific activity that can be interpreted within the context of known function. Thus, ferritin, containing iron ionic clusters, is a very likely candidate to protect DNA from oxidation. Hydrolase activity might be necessary in the phosphonate degradation. Methylation is known to be important in synchronizing the cell division process. The reduction of dangerous oxidating agents may involve mechanisms similar to that of peroxide reduction. In all such cases, analogies such as these could not be treated as proof of a prediction. While they indirectly support the prediction, the final verification must be done by experiment.

Predictions for hypothetical proteins

For hypothetical proteins, verification of the structural predictions is even more difficult, because nothing is known about a prediction target. On the other hand, if the analysis of the model supports possible functional similarity between target and template, a functional prediction might be attempted (Table 3).

1. The structure of the hypothetical protein YJDC is predicted to be similar to that of the tetracycline repressor 2tct. The structure and function of this protein are unknown, but it is interesting to note that this protein is predicted to be similar to a large family of proteins, which includes some hypothetical operon repressors (ACRR, ENVR, and UIDR), transcriptional regulators (YCDC), and several hypothetical proteins. All the proteins from this family with known functions are involved in regulating resistance to antibiotics and toxic hydrophobic substances.

Table 3. *A sample from about 66 high significance predictions for hypothetical proteins from the E. coli genome*[a]

E-value	PDB code and name of a template		Target	BLAST prediction		
0.05	1fbaA	Aldolase	YIHT		>10	
0.05	2tct_	Tetracycline repressor	YJDC		>10	
0.03	1ukz_	Uridylate kinase	YACE		>10	
0.01	1din_	Dienelactone hydrolase	YIEL	2lip	4.6	Y
9e–3	1occA	Cytochrome *c* oxidase	YICE		>10	
6e–3	1ctt_	Cytidine deaminase	YFHC		>10	
5e–3	1din_	Dienelactone hydrolase	YEIG	2lip	0.3	Y
5e–3	2pia_	Dioxygenase reductase	f254		>10	
4e–3	1sxl_	Sex-lethal protein	YJAI		>10	
1e–4	2tysA	Tryptophan synthase	YJCU	1qap	0.2	N

[a] SwissProt codes for hypothetical proteins are given if available.

2. The hypothetical protein YIHT is predicted to have a structure similar to fructose-1,6-biphosphate aldolase. YIHT is closely homologous to tagatose aldolase, and PSI BLAST finds a marginal hit (E-value 0.60) to a fructose aldolase from *Onchocerca vulvulus* (filarial worm).

3. Hypothetical protein YACE is predicted to be similar to uridilate kinase. This prediction is supported by a weak similarity (E-value of 0.1) to shikimate kinase from blue-green algae recognized by PSI BLAST. In this example, as in the suggested homology between gluconokinase, shikimate kinases, and uridilate kinases discussed earlier, the similarity is much stronger in the ATP binding domain than it is in the catalytic domain. All residues involved in the ATP binding are conserved in all models. At the same time, no active site residues can be identified in the alignments and, thus, in the model. A model of YACE protein is shown in Figure 2. Conserved residues in the ATP binding domain, forming a classical mononucleotide binding motif with the strongly conserved G-X-X-G-X-G-K sequence (Schulz, 1992), are shown in atomic detail. On the other hand, five hydrophobic residues replace the five positively charged residues involved in the NMP binding site in uridilate kinase. Note that the model of the catalytic domain displays a large unstructured loop, characteristic of a very low quality model. The quality of this part of the model, as measured by threading, is also very low.

4. Hypothetical protein YHFC is predicted to be similar to cytidine deaminase. This prediction is supported by a strong homology between YHFC and two other families of deaminases: riboflavin specific and cytosine deaminases. Here, the target protein is shorter than the template and aligns to only one domain of the template. A model of YACE protein is shown in Figure 3. Residues involved in zinc binding are conserved and are shown in atomic detail on the model.

5. Hypothetical protein YJAI is predicted to be similar to the sex lethal protein. This prediction is supported by a weak homology (E-value of 0.02) to nucleolin, another RNA binding protein, containing an RNP motif, characteristic of DNA/RNA binding proteins, and present in the sex lethal protein. However, the YJAI itself does not have the RNP motif.

A detailed analysis of the models of the hypothetical proteins discussed above supports the hypothesis that some of the elements of the new proteins and their functions might be related to those of their homologous counterparts.

Insights from structural predictions

In many cases, the recognition of very distant homology enables us to realize that different protein families and their activities, which were thought to be unrelated, are actually homologous. In many

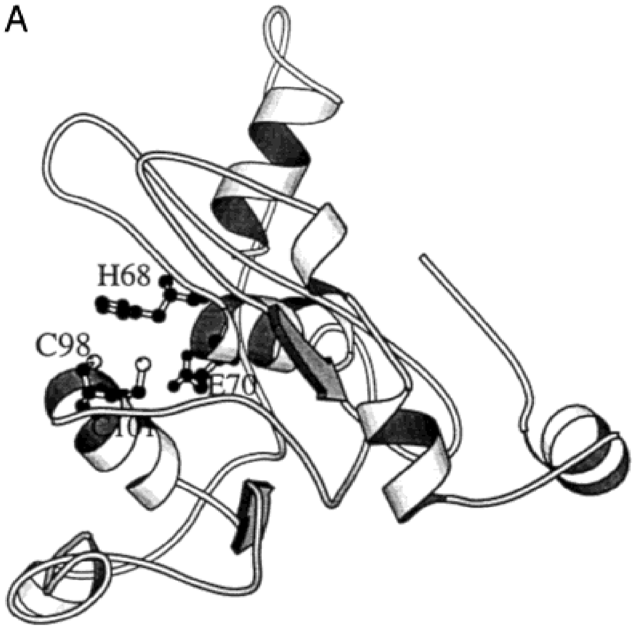


Fig. 2. A: A model of the hypothetical protein YACE from *E. coli*, built on the uridilate kinase template (PDB code 1ukz). ATP binding residues are shown in atomic detail. **B:** The alignment obtained from the BASIC algorithm, with ATP binding residues highlighted both in the template and in the model sequence. The same alignment was used to build the model shown in **A**.

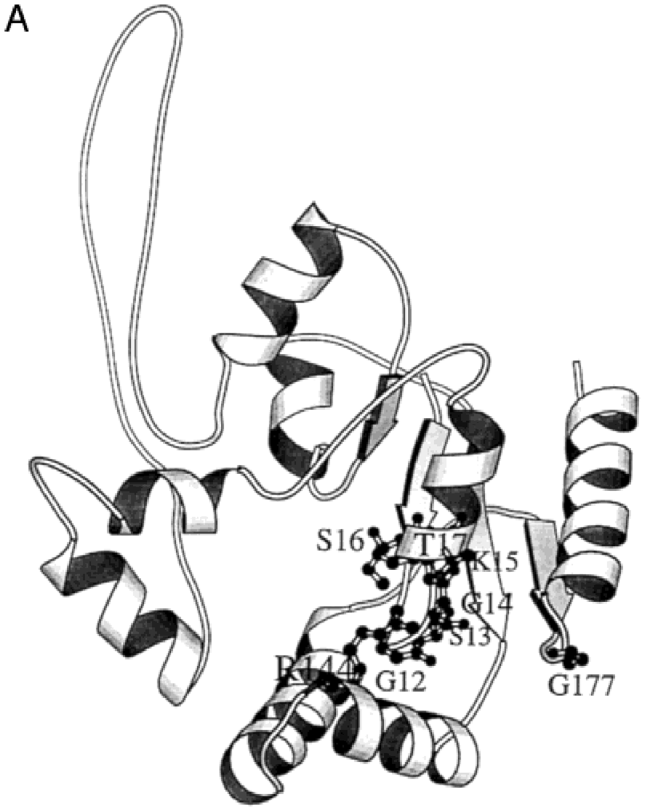
B

```
1ukz_        PAFSPDQVS- --VIFVLGGP GAGKGTQCEK LVKDYSFVHL SAGDLLRAEQ GRAGSQYGEL
YACE_ECOLI   ---------M RYIVALTGGI GSGKSTVANA FAD-LGINVI DA-DIIARQV VEPGAPALHA

1ukz_        IKNCIKEGQI VPQ------- ---------- ------EITL ALLRNAISDN VKANKHKFLI
YACE_ECOLI   IADHFGANMI AADGTLQRRA LRERIFANPE EKNWLNALLH PLIQQETQHQ IQQATSPYVL

1ukz_        DGFPRKMDQA ISFERDIVES -KFILFFDAP EDIMLERLLE RGKTSGRSDD NIESIKKRFN
YACE_ECOLI   ------WVVP LLVENSLYKK ANRVLVVDVS PETQLKRTMQ RDDV------ TREHVEQILA

1ukz_        TFKETSMPVI EYFETKSKVV RVRADRSVED VYKDVQDAIR DSL------- -----
YACE_ECOLI   AQATREA--- ---RLAVADD VIDNNGAPDA IASDVARLHA HYLQLASQFV SQEKP
```



B

```
1ukz_          PAFSPDQVS- --VIFVLGGP GAGKGTQCEK LVKDYSFVHL SAGDLLRAEQ GRAGSQYGEL
YACE_ECOLI     ---------M RYIVALTGGI GSGKSTVANA FAD-LGINVI DA-DIIARQV VEPGAPALHA

1ukz_          IKNCIKEGQI VPQ------- ---------- ------EITL ALLRNAISDN VKANKHKFLI
YACE_ECOLI     IADHFGANMI AADGTLQRRA LRERIFANPE EKNWLNALLH PLIQQETQHQ IQQATSPYVL

1ukz_          DGFPRKMDQA ISFERDIVES -KFILFFDAP EDIMLERLLE RGKTSGRSDD NIESIKKRFN
YACE_ECOLI     ------WVVP LLVENSLYKK ANRVLVVDVS PETQLKRTMQ RDDV------ TREHVEQILA

1ukz_          TFKETSMPVI EYFETKSKVV RVRADRSVED VYKDVQDAIR DSL------- -----
YACE_ECOLI     AQATREA--- ---RLAVADD VIDNNGAPDA IASDVARLHA HYLQLASQFV SQEKP
```

Fig. 3. A: A model of the hypothetical protein YHFC from *E. coli*, built on the cysteine deaminase template (PDB code 1ctt). The active site residues are shown in atomic detail. **B:** The alignment obtained from the BASIC algorithm, with the active site residues highlighted both in the template and in the model sequence. The same alignment was used to build the model shown in **A**. Note that only one domain of 1ctt was used for modeling.

cases, the importance of these findings extends beyond a simple structure prediction of a single protein. In addition to making a structural prediction, they point to undiscovered pathways or make connections between apparently unrelated functions.

For instance, several hypothetical proteins from the *E. coli* genome show very strong similarity to the structural family containing bromoperoxidase (1bro) and lipase (1tah). At the same time, these proteins show sequence similarity to proteins from other bacteria known to be involved in polyhydroxybutyrate metabolism.

Polyhydroxybutyrate (PHB) is a polymer produced by several bacteria as energy storage material. Its physical properties, which are similar to that of polyethylene, coupled with the easy biodegradability of PHB, make it an attractive target for industrial application (Anderson & Dawes, 1990). To study the feasibility of the commercial biosynthesis production of PHB, polyhydroxyalkanoate synthesis genes from *Alcaligenes eutrophus* were introduced into the *E. coli* genome (Wang & Lee, 1997). Wild-type *E. coli* does not use PHB as energy storage; therefore, it was argued, it does not contain PHB synthesis and degradation apparatus. This hypothesis was strengthened by the observation that no homologues of PHB related proteins were identified in the *E. coli* genome. Recently, however, the low molecular weight PHB analogue was found in *E. coli* in small amounts under growth limiting or induced genetic competence conditions (Huang & Reusch, 1996). At the same time, PHB and/or its analogues were identified to be much more ubiquitous than previously thought. Low molecular weight complexes of PHB were discovered to form membrane spanning channels, conducting DNA, ions, and other substances. Therefore, *E. coli* and other organisms must have enzymes involved in PHB metabolism, even though they were not identified as yet.

Analysis of the known PHB polymerases and depolymerases sequence yielded some interesting regularity. All PHB depolymerases contain a characteristic GxSxG pattern, which is a part of the active site. At the same time, all polymerases have a GxCxG pattern in analogous position. Using two criteria, structural similarity as predicted by BASIC and conservation of the characteristic active site pattern, it was possible to identify 12 potential PHB depolymerases and three potential PHB polymerases. With the exception of one protein, all of the 15 identified proteins had no identified function. It is likely that at least some of these uncharacterized proteins take part in the PHB metabolic pathway. The complete list of predictions in both groups is given in Table 4. One of the most interesting proteins in this group is protein coded by gene f136 (Fig. 4). Despite its marginal significance as a potential polymerase (E-value of 0.6), its model has a complete active site with a C-H-D triad, typical for PHB polymerases. At the same time, this protein is very small and the putative active site is on the protein surface. One can speculate that this unusual position of the active site might be connected to the PHB role as a tightly wound transmembrane channel.

Discussion

Fold assignments for the entire *E. coli* genome were carried out using the position specific iterative BLAST algorithm (Altschul et al., 1997) and a new BASIC algorithm (Rychlewski et al., 1998). For almost 30% of all *E. coli* proteins, there is a very high probability that their structure is similar to that seen in one of the already characterized protein superfamilies. For an additional 30% of *E. coli* proteins, predictions could be made with lower significance and more incorrect predictions are expected to be in this group. The database of prediction results is available at cape6.scripps.edu. Fold prediction by recognition of distant homologies is part of a general problem annotating newly sequenced proteins by comparing them to already known and characterized proteins. Methods described here, as well as methods developed in other groups, change the perspective on analysis of newly sequenced proteins, such as those methods from genome projects. A

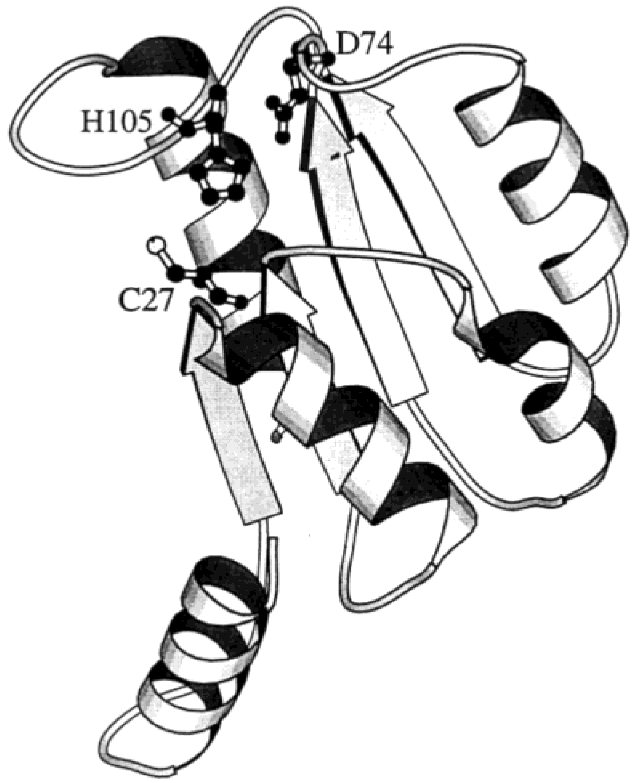


Fig. 4. A model for a hypothetical protein f136 (GenBank accession number 1789373), built on the lipase template. Note that only one domain of the template was used in modeling. The active site residues are shown in atomic detail.

Table 4.

E-value[a]	Zscore[b]	1 3 2[c]	*E. coli*[d]	Gen[e]	Description[f]
					Potential depolymerases
0.0	45.6	S H D	o309	1786545	MHPC_ECOLI 2-HYDROXY-6-KETNONA-2,4-DIENEDIOIC ACID HYDROLASE
0.0	43.3	S H S	f254	1786902	
0.0	40.7	S H D	bioH	1789817	BIOH_ECOLI BIOH PROTEIN
0.0	37.2	S H D	yfbB	1788598	YFBB_ECOLI HYPOTHETICAL 26.7 KD PROTEIN IN MEND-MENB
0.0	30.5	S H D	yheT	1789752	YHET_ECOLI HYPOTHETICAL 38.5 KD PROTEIN IN KIFB-PRKB
0.0	29.8	S H D	yjfP	1790634	YJFP_ECOLI HYPOTHETICAL 27.6 KD PROTEIN IN AIDB-RPSF
6.2E−07	22.7	S H D	o293	1788884	
1.1E−06	22.0	S H D	yeiG	1788477	YEIG_ECOLI HYPOTHETICAL 31.3 KD PROTEIN IN FOLE-CIRA
1.2E−06	22.1	S H D	f277	1786551	YAIM_ECOLI HYPOTHETICAL 31.4 KD PROTEIN IN MHPT-ADHC
1.5E−05	19.5	S H D	yieL	1790156	YIEL_ECOLI HYPOTHETICAL 44.1 KD PROTEIN IN TNAB-BGLB
0.156	10.2	S H D	f240	1788817	YPFH_ECOLI HYPOTHETICAL 25.7 KD PROTEIN IN DAPE-PURC
0.276	9.6	S H D	ybaC	1786682	YBAC_ECOLI HYPOTHETICAL 36.0 KD PROTEIN IN HEMH-GSK
					Potential polymerases
4.1E−0.5	18.5	C H A	f310	1787587	
9.8E−03	13.0	C H A	f332	2367305	
4.65	6.8	C H D	f136	1789373	

[a] E-Value is the number of random hits that could be obtained with this Z-score.
[b] Zscore is the Z-score of the highest similarity to any of the proteins from the PHB-metabolizing group.
[c] 1 3 2 are residues aligned to active site residues (S-H-D) or (C-H-D).
[d] *E. coli* is a general description.
[e] Gen is the general number.
[f] Description is the SwissProt description of the *E. coli* protein.

large percentage of proteins coded by genomes could be assigned a three-dimensional fold, allowing to use detailed understanding of function, which comes from knowing the protein structure, to analyze and study these proteins.

The final step in the fold prediction, building the full three-dimensional models of all predicted structures by automated modeling, could be used for prediction verification. It is important to note that models built by automated modeling do not represent the best possible models that could be built from the given template. Models prepared by a human expert and undergoing several iterations of improvements would be much closer to the real structure than the models prepared for this analysis. However, such modeling requires a substantial investment of time and expertise that is appropriate only in detailed studies of specific systems. On the other hand, automated modeling makes it possible to introduce full atom models for genome-scale studies. Model quality in automated modeling is worse than in studies done by human experts, but as illustrated earlier on the example of cysR, it is sufficient to represent a useful first approximation, particularly appropriate for first attempts at the functional prediction.

Only a small, randomly chosen group of predictions were analyzed in depth in this paper. These and all other predictions are made public with an invitation to verify or refute them by experiment or with other prediction algorithms. The evaluation of bone fide predictions is difficult at the time they are made. For a large number of the *E. coli* proteins, their function is known and can be checked for compatibility with the structural prediction. The mutual position of catalytic residues, agreement between known functional features, and the predicted structure and functional analogies between the structural superfamily and its new predicted member can be used to argue for or against the prediction being correct. Recently, an automated procedure for analysis of active site residues was implemented and applied for an automatic search for proteins with glutaredoxin/thioredoxin activity (Fetrow & Skolnick, 1998). Such automated function verification is a perfect tool for a genome scale protein function prediction and the results of such an analysis on the protein structural predictions from the *E. coli* genome are presented in Fetrow et al. (1998).

Structure prediction is only a prelude to a much more interesting and important prediction of function. In recent years, there has been a growing consensus that most proteins with similar structures and apparently dissimilar sequences are actually homologous. This increases the importance of structure prediction, because it is not just the structure that is being predicted, but actually a relationship between an uncharacterized protein and an already well-studied family. Similar structure is only one of the many features that could be shared between various members of the family. Zhang et al. (1999) present tools for automated evaluation of function conservation in homologous families, bringing us a step closer to automated function prediction.

Several detailed examples presented here illustrate benefits of having a fold prediction. A most interesting situation arises when a general function of a protein is known, but a specific activity is not. In such cases, structural prediction, if accompanied by conservation of active site residues, can provide a specific prediction of how a general function is carried out. For instance, in the case of the DPS protein, the general function of protecting DNA from starvation induced oxidation was known from genomic experiments. Prediction of its homology to ferrodoxin provided a detailed prediction about its mechanism.

Methods

PSI BLAST and the sequence database

The position specific iterative BLAST algorithm (Altschul et al., 1997) is the newest version of the de facto standard of database protein similarity searching algorithms. This algorithm addresses the principal shortcoming of the previous BLAST algorithm: its inability to introduce gaps in the alignment. In addition, the PSI BLAST algorithm allows the iterative building of a sequence profile from the multiple alignment of sequences of homologous protein identified in the first pass of the algorithm. The PSI BLAST program was downloaded from the NIH web site and used following the guidelines in the manual. Specifically, in the application described here, the gapped BLAST step was followed by one iteration of PSI BLAST based on the profile created from proteins identified in the first run with E-values better that 0.1. On tests with the fold recognition benchmark, we have found that this strategy produces the best recognition ratio.

The sequence database used by the PSI BLAST algorithm contains a nonredundant compilation of sequences available from SWISSPROT and PIR databases, as well as translated DNA sequences from EMBL and NCBI nucleotide sequence databases and sequences of all proteins deposited in the Brookhaven PDB. This database was used to prepare sequence profiles for all targets and templates and is a complete superset of the database used by the BASIC method. The version used in this work was compiled in November 1997.

Profile sequence preparation

The method described in this paper is based on an evaluation of the similarity between two sequence profiles. A sequence profile is a position specific probability distribution, which for every position along the sequence gives a probability that one of the 20 amino acids would occupy this position (Gribskov et al., 1987; Bork & Gibson, 1996). Profiles were generated automatically using the multiple alignment of homologous sequences as generated by the PSI BLAST algorithm. Exactly the same procedure is followed for the target proteins as for all proteins contained in the databases being searched.

Databases of sequence profiles

The database of 1,151 representative protein structures was prepared on the basis of a nonredundant set of protein structures included in the FSSP database as available from the DALI server at EBI (DALI, 1995). A version of the DALI sets from January 1998 was used. The exact list used for this work is available at cape6.scripps.edu. To avoid possible differences between databases used by BLAST and BASIC algorithms, sequences of all proteins from the DALI set were added again to the large sequence database and identified with the PDB keyword and their PDB code name.

The BASIC profile-to-profile alignment algorithm

Two sequence profiles are compared in the same way as two sequences. A local-local version of a Smith–Waterman dynamic programming algorithm is used (Waterman, 1995). The similarity

Table 5. *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	—	—
Basic THREADING	22	30	34
Global sequence alignment	40	50	52
Hybrid THREADING	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 sometimes are not listed, due to a large number of homologous proteins.

score between positions in two sequences is calculated with the mutation matrix, such as the Gonnett similarity matrix (Gonet et al., 1992). For two profiles, this value is calculated as an average of scores between all amino acid pairs, averaged according to the probability distribution in each profile. Three parameters, gap introduction penalty, gap extension penalty, and a constant added to each element of the mutation matrix are optimized for a fold recognition benchmark (Rychlewski et al., 1998).

Optimization and verification of the BASIC algorithm

The BASIC algorithm was optimized to recognize the maximum number of structurally similar proteins on benchmarks customized for fold prediction algorithms. A particular benchmark available from the web server at University of California, Los Angeles (UCLA) 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 5) show that a sequence-only fold recognition method can closely match the prediction accuracy of the best threading algorithms.

Prediction significance

Scores of individual profile–profile comparisons are corrected for the size of proteins being compared (Karlin & Altschul, 1990; Waterman, 1995) and used to calculate the distribution of scores for a given prediction target. The empirical distribution was fitted to an extreme value distribution. The parameters of this fit were used to calculate the E-value, i.e., the expected number of proteins with a given score in a given database.

The estimation of the reliability of the prediction was based on the E-value statistic. The cutoff of the 0.05 E-value used here is much larger than the scores of the false positive answers of the procedure observed during the development. The biggest E-value for a false positive in the UCLA benchmark (Rychlewski et al., 1998) was equal to 1.6. However, it is not known how different the distribution of scores on the training set is from the distribution on the larger set used in the actual predictions. For this reason, we use a very conservative significance threshold.

Automatic model building and evaluation

Alignments between each target and the best scoring template were prepared by the BASIC algorithm. These alignments were reformatted and used as input to the MODELLER program (Sali & Overington, 1994). This program builds full three-dimensional models of target proteins using the method of "satisfaction of spatial restraints" (Sali & Overington, 1994). A standard MODELLER routine "model" was applied.

It is important to note that in this application only a small subset of MODELLER potential was used. Automatic modeling is inferior to that done by human experts, but in genome scale structural predictions it is difficult to relay on human expertise.

The alignments and resulting models were analyzed for quality. At first, the alignments were analyzed in conjunction with the template structure. Long insertions or deletions in the core of the protein, which would be difficult to accommodate without large global rearrangements, could be identified before the actual modeling step. Models with excessive problems of this type were unlikely to be correct. The second step involved an analysis of the quality of the complete model. Steric overlaps and wrong bond geometries were checked with ProCheck and the overall model quality with threading energy. Unfortunately, in such an analysis it is almost impossible to distinguish between template errors (i.e., wrong fold prediction) and alignment errors (i.e., correct global fold but wrong local details). Therefore, a continuous, well-packed, and low energy model is a strong indication of a correct prediction, but an obviously wrong, low quality model does not necessarily indicate a wrong fold prediction.

Function prediction and its evaluation

Recognition of a possible homology could be automatically interpreted as a function prediction. In its simplest form, the prediction is that the function of a new protein is the same as its putative homologue. This prediction can be checked by analysis of the conservation of residues involved in the biochemical activity of a protein with a known structure. Two public databases containing information about such residues, PDBsum (Laskowski et al., 1997) and PROSITE (Bairoch, 1994), were used to define "function signatures." PDBsum is the database of summaries of information about structure from PDB, available from the UCLA web site (Laskowski et al., 1997). PROSITE is a dictionary of protein sites and patterns, developed at the University of Geneva (Bairoch, 1994). At this step, residues mentioned in either of the two databases as important for function were identified and their conservation checked in the alignment.

Program and prediction database availability

A version of the BASIC program is available at cape6.scripps.edu. It offers the possibility of similarity predictions in the database of structural families. The user can supply the sequence of the target protein and a fold assignment and a full structural model is returned.

Note added in proof

The cape6.scripps.edu server is now being moved to bioinformetics.burnham-inst.org.

Acknowledgments

The authors are grateful for many stimulating discussions with Drs. A. Kolinski, J. Skolnick, and J. Fetrow. We also thank Dr. A. Sali for providing us with a MODELLER software package, as well as for help and advice in using it. Ribbon diagrams were prepared using the MOLSCRIPT program, kindly provided by Dr. P. Kraulis. This work was supported by NIH Grant No. GM48835.

References

Alexandrov NN, Nussinov R, Zimmer RM. 1996. Fast fold recognition via sequence to structure alignment and contact capacity potentials. *Pacific symposium on biocomputing*, Volume 96. Hawaii: World Scientific.

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. *Nucl Acid Res 25*:3389–3402.

Anderson AJ, Dawes EA. 1990. Occurrence, metabolism, metabolic role and industrial uses of bacterial polyhydroxyalkanoates. *Microbiol Rev 54*:450–472.

Bairoch A. 1991. PROSITE: Dictionary of protein sites and patterns. *Nucleic Acid Res 19S*:2241–2245.

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.

Bozzi M, Mignogna G, Stefanni S, Barra D, Longhi C, Valenti P, Chiancone E. 1997. A novel nonheme iron-binding ferritin related to the DNA-binding proteins of the Dps family in *Listeria innocua*. *J Biol Chem 272*:3259–3265.

Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through folding motif. *Proteins 16*:92–112.

Casari G, Ouzounis 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.

DALI. 1995. *Protein structure comparison by alignment of distance matrices.* Heidelberg: EMBL.

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 oxireductase activity. *J Mol Biol*. Forthcoming.

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 282*:703–711.

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. *Nature Str 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.

Gonnet GH, Cohen MA, Benner SA. 1992. Analysis of amino acid substitution during divergent evolution. *Science 256*:1443–1445.

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

Huang R, Reusch RN. 1996. PHB is associated with specific proteins in the cytoplasm and membranes of *E. coli*. *J Biol Chem 271*:22196–22202.

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

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

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.

Matsuo Y, Nishikawa K. 1994. Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Sci 3*:2055–2063.

Moxon ER, Higgins CF. 1997. A blueprint for life. *Nature 389*:120–121.

Orengo CA, Flores TP, Jones DT, Taylor WR, Thornton JM. 1993. Recurring structural motifs in proteins with different functions. *Curr Biol 3*:131–139.

Ouzounis C, Sander C, Scharf M, Schneider R. 1993. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from three-dimensional structures. *J Mol Biol 232*: 805–825.

Quiocho FA, Ledvina PS. 1996. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: Variations of common themes. *Mol Microbiol 20*:17–25.

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.

Sali A. 1994. MODELLER. A protein structure modeling program. Available from http://guitar.rockefeller.edu.

Sali A, Overington JP. 1994. Derivation of rules for comparative modeling from a database of structural alignments. *Protein Sci 3*:1582–1596.

Schulz GE. 1992. Binding of nucleotides by proteins. *Curr Opin Struct Biol 2*:61–67.

SCOP. 1995. *Structural classification of proteins*. MRC Cambridge.

Tyrrell R, Vershueren KHG, Dodson EJ, Murshudov GN, Addy C, Wilkinson AJ. 1997. The structure of the cofactor binding fragment of the LysR family member, CysB: A familiar fold with a surprising subunit arrangement. *Structure 5*:1017–1032.

Wang F, Lee SY. 1997. Production of PHB by fed-batch culture of filamentation-suppressed recombinant *E. coli*. *Appl Environ Microbiol 63*:4765–4769.

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

Wilmanns M, Eisenberg D. 1995. Inverse protein folding by the residue pair preference profile method. *Protein Eng 8*:626–639.

Yi TM, Lander ES. 1994. Recognition of related proteins by iterative template refinements. *Protein Sci 3*:1315–1328.

Zhang B, Rychlewski L, Pawlowski K, Fetrow J, Skolnick J, Godzik A. 1999. Flow fold predictions to function predictions. *Protein Sci*. Forthcoming.