

Microbial Genomes Have Over 72% Structure Assignment by the Threading Algorithm PROSPECTOR_Q

Daisuke Kihara and Jeffrey Skolnick*

UB Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York

ABSTRACT The genome scale threading of five complete microbial genomes is revisited using our state-of-the-art threading algorithm, PROSPECTOR_Q. Considering that structure assignment to an ORF could be useful for predicting biochemical function as well as for analyzing pathways, it is important to assess the current status of genome scale threading. The fraction of ORFs to which we could assign protein structures with a reasonably good confidence level to each genome sequences is over 72%, which is significantly higher than earlier studies. Using the assigned structures, we have predicted the function of several ORFs through “single-function” template structures, obtained from an analysis of the relationship between protein fold and function. The fold distribution of the genomes and the effect of the number of homologous sequences on structure assignment are also discussed. *Proteins* 2004;55:464–473. © 2004 Wiley-Liss, Inc.

Key words: threading; genome-scale protein structure prediction; fold distribution; protein function prediction

INTRODUCTION

With the recent completion of a number of genome sequencing projects,¹ a key goal is to identify the function of all the open reading frames (ORFs) in a given genome. Partly to aid in this goal and partly to elucidate the nature of protein structure space, various structural genomics projects² have embarked on determining the tertiary structure of a significant fraction of a number of genomes. If the goal is to make this structure determination process as efficient as possible in determining novel folds, then it is important to identify those proteins whose domains have structures similar to already known structures. Furthermore, in a series of papers, we have demonstrated that even if the resulting predicted structures are of low resolution, say having a root mean square deviation (RMSD) from native, for the α -carbons below 6 Å, such structures can often be used to predict the enzymatic function of a protein.^{3–5} Thus, it is worthwhile to perform genome scale tertiary structure prediction^{6–8} and even more recently quaternary structure predictions,^{9–11} and a number of workers have been engaged in this task.^{12,13}

Before delving into applications of threading to entire genomes, a brief overview of the state of the art is appropriate. To assess whether the fold of a new protein sequence, the target, has previously been solved, (i.e.,

matches a known template structure), two types of approaches have been developed. Roughly speaking, these divide into sequence comparison and structure-based methods. Sequence-based approaches are designed to establish whether an evolutionary relationship between two protein sequences exists. If so, because fold tends to be better conserved than function, one can then identify the fold of the target. The most powerful of the recently developed sequence-based approaches tend to be iterative and aim to construct a sequence conservation profile by pooling sequences identified on successive iterations; a prototypical example is PSI-BLAST.¹⁴ More recent innovations include profile-profile comparisons that compare the sequence conservation profile of the target with that of the corresponding profile of the template sequences.¹⁵ Hidden Markov Models¹⁶ (HMMs) represent yet another powerful class of sequence based algorithms. These include Pfam¹⁷ and SAM-T99.¹⁸ Alternatively, although many of the most successful of the threading algorithms have a strong evolutionary component, threading includes structural information into the fold assignment process. As convincingly demonstrated in CASP5, there are a number of such approaches that now significantly outperform PSI-BLAST.¹⁴ In this respect, this paper describes the application to genome scale tertiary structure prediction of the latest version of threading algorithm, PROSPECTOR_Q, an earlier variant (PROSPECTOR 1.0),¹⁹ which was a key contributor to our reasonably successful performance in CASP5.²⁰

In the recent past, there have been a number of both threading and sequence-based whole genome analyses. Despite these earlier reports, we have revisited this issue because structure assignment to ORFs can often provide significant insights into ORFs' function, ligand interactions, and possible role in the pathways. As shown in CASP5, the performance of threading methods has been improving both because of the increasing sophistication of the algorithms and the expansion of structure/sequence databases. Thus we believe it is important to ascertain the current status of genome-scale protein structure predic-

Grant sponsor: the National Institutes of Health, Division of General Medical Sciences; Grant number: GM-48835; Grant sponsor: the Oishei Foundation.

*Correspondence to: Jeffrey Skolnick, Center of Excellence in Bioinformatics, University at Buffalo, 910 Washington Street, Suite 300, Buffalo, NY 14215. E-mail: skolnick@buffalo.edu

Received 25 June 2003; Accepted 10 October 2003

Published online 5 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20044

tion. We have selected the following five microbial genome sequences to demonstrate the performance of our threading method, *Mycoplasma genitalium*, *Escherichia coli*, *Bacillus subtilis*, *Aquifex aeolicus*, and *Saccharomyces cerevisiae*. These genomes were frequently used in prior studies,^{21–28} especially *Mycoplasma genitalium*, so that we can compare our results to them. Our results are typical of those from over a dozen genomes including human that will be presented in subsequent papers.

This paper is organized as follows: First, we briefly describe our threading algorithm PROSPECTOR_Q and then present the results on a large-scale benchmark set. The details can be found elsewhere.²⁹ Then, an analysis is made of the extent of fold assignment to five genomes, as well the percent of residues in a genome assigned to structures. Comparison is made to results obtained from FASTA,³⁰ PSI-BLAST,¹⁴ PEDANT,³¹ GTOP³² and other earlier published works.^{21–28} Next, the predicted protein fold distribution for the five genomes is shown. Finally, the ability to use fold to uniquely assign protein function is investigated. Here, we have analyzed protein structure and function relationship as the correspondence between the EC numbers and CATH fold,³³ to select “single-function” template structures, which were used for function assignment. Our protein structure assignments to the genome sequences can be found at <http://www.bioinformatics.buffalo.edu/resources/genomethreading/>.

MATERIALS AND METHODS

Overview of PROSPECTOR_Q

Here, we briefly describe PROSPECTOR_Q. See Skolnick and Kihara¹⁹ and Skolnick et al.²⁰ for additional details. PROSPECTOR_Q takes two kinds of multiple sequence alignments (MSAs) to the target sequence as input: close homologous sequences (35 to 90% pairwise sequence identity) and sequences whose FASTA³⁰ E-values are less than ten to the target sequences. The former is used to generate a “close sequence profile,” and the latter to generate a “distant sequence profile.” The resulting sequence profile alignment generates the partners to be used in the evaluation of the pair interactions expressed as the average over all close or distant sequences of a quasi-chemical residue-residue pair interaction potential (hence the **Q** in PROSPECTOR_Q). Also included is a secondary structure prediction term. Next, consensus residue contacts among top hit template proteins are converted into a protein-specific pair potential³⁴ that is combined with the aforementioned homology averaged quasi-chemical pair potential and used in the next iteration. This process is done a total of three times. A new feature of PROSPECTOR_Q is that the score is evaluated as the energy difference between the best score of the target sequence aligned to the template and the reversed sequence aligned to the template, following the idea of Karplus;³⁵ a comparison with sequence randomization which is much more expensive is summarized below. We also reduced the gap penalties at the beginning and end of the alignment to enhance the fold recognition ability of the method.

The threading template library consists of a representative subset of proteins in the PDB,³⁶ such that no pair of

template proteins has more than 35% pairwise sequence identity calculated over either the aligned region, or over the smaller of the pair of template proteins. As of February 2003, there are 3595 such proteins.

As shown elsewhere²⁹ a benchmark was designed to identify templates with little if any apparent homology to the target sequence. For proteins below 200 residues, there are 1491 such target sequences that are no more than 35% identical to each other, with no more than 30% identical to any template. Of these cases, 1109 (75%) have identified templates whose Z-score (the energy in standard deviation units relative to the average) is greater than 7.0 and at least a 20-residue-long alignment. The average global sequence identity is about 22%. About 95% of the target sequences have a good structural alignment for the best scoring template, with an average coverage of 78% and an average root mean square deviation, RMSD, from native of the C α atoms of 2.95 Å. Thus, even if only one template is identified with a Z-score > 7, there is a very high likelihood that the fold of the target protein has been correctly predicted. We now turn to the question of the absolute accuracy of the predicted alignment, not just fold assignment. If the best one is identified among the top five scoring templates, with an average rank of 1.3 (1 means that only the top template need be selected), then 73% of the target sequences (with at least one template with Z-score > 7) have at least one template alignment below 6.5 Å (a reasonable cutoff for structural similarity), with an average RMSD from native of 3.7 Å and 82% coverage for all aligned residues. Moreover, 886/1109 (80%) have a very good alignment over a significant fraction of their structure (with a local RMSD no more than 5 Å) and an average RMSD of 2.3 Å with 71% coverage. If only the highest Z-score template is used, then 57% have an RMSD below 6.5 Å, with an average RMSD of 3.5 Å for all aligned residues with a coverage of 86%. Seventy percent (70%) of the highest Z-score templates have a significant fragment below 5 Å, with an average RMSD of 2.0 Å and 71% coverage. This is the typical expected accuracy for the nontrivial cases of low sequence identity to templates that we might expect to encounter when PROSPECTOR_Q is applied to entire genomes.

If we compare the results of PROSPECTOR_Q when the Z-score is calculated relative to randomized sequences (20 randomizations was used), we find the following: Sequence randomization finds 1061 targets with at least one template having a Z-score > 7. Nine hundred and ninety-eight (998, 94%) of these targets are the same as when sequence reversal is used. Using the highest Z-scoring template from sequence randomization, 947 have the same template as identified using sequence reversal, with average rank 1.2. There are 813 (76%) targets with a RMSD < 6.5 Å, an average RMSD of 3.8 Å and 81% coverage. Of these, 800/813 are the same targets as when sequence reversal is used; 89% have the same template as rank 1. If we consider those targets having well identified regions, then sequence reversal provides good targets in 80% of the cases, while randomization is about 4% better. Given the additional cost (a factor of 20 in computer time), the

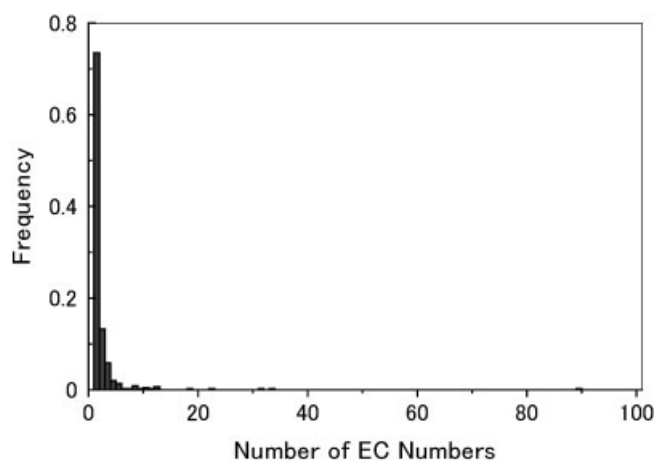


Fig. 1. Histogram of CATH topologies associated with a given number of different EC numbers. Only the first three digits of the EC numbers are counted.

improvement using sequence randomization is hard to justify, with essentially identical results obtained.

Genome Sequences

The sequences of the ORFs of the five analyzed organisms are taken from the KEGG database (<ftp.genome.ad.jp/pub/kegg/genomes/genes>).¹ To build the sequence profiles, we constructed a sequence database by combining Swiss-Prot (<http://www.expasy.ch/sprot/>)³⁷ and the KEGG genome sequence database. For each ORF, we have prepared the close and the distant profiles. FASTA^{38,39} was used to select homologous sequences from the sequence database, and then clustalW⁴⁰ was used to generate the MSAs.

Identification of Single-Function Topologies

For each protein topology classified in CATH database³³ (i.e., the first three levels/digits in the classification scheme), the associated EC numbers of the protein structures (i.e., PDB entries) are identified. For a protein structure, EC numbers are extracted not only from the PDB file itself, but also collected through related Swiss-Prot³⁷ and Enzyme database⁴¹ entries using the BioMolQuest system.⁴² Figure 1 shows a histogram of the CATH topologies with the associated number of the EC numbers (the first three levels/digits are counted). In this analysis, we have discarded topologies classified into classes 5–9 in CATH, because these classifications are preliminary (i.e., only the CATH topologies whose first digit is 1–4 are counted). As shown in Figure 1, besides several multi-functional topologies, such as TIM barrels or ferredoxins, more than 70% of the topologies are associated with only one EC number. This procedure constitutes a preliminary screening of “single-function” topologies, or more precisely, “single-enzyme-function” topologies, since the EC number is used to classify protein function. There are 260 topologies (represented by 410 threading template proteins in our template database) that are selected. On average, there are three protein sequences associated with each topology in CATH (at the 90% sequence homology level).

RESULTS

We have applied PROSPECTOR_Q to five representative complete genome sequences, namely, *Mycoplasma genitalium*, *Escherichia coli*, *Aquifex aeolicus*, *Bacillus subtilis*, and *Saccharomyces cerevisiae*. Summaries of the fold assignments are shown in Table I. For each genome, structures are assigned to more than 72% of the ORFs, with the highest assignment of 85.2% belonging to *A. aeolicus*. The Z-score threshold used here was 7.0, which is the same as the one used in the benchmark described in the previous section. For comparison, we have employed FASTA³⁰ and PSI-BLAST¹⁴ (the fourth to sixth column from the right in Table I), because these are two major programs for homology searches, and also show results from other sources: GTOP,³² PEDANT,⁴³ and results from Gerstein’s group done in 2000. In the sixth and fifth column, FASTA and PSI-BLAST were simply run against protein sequences in the PDB.³⁶ The E-value threshold used for the FASTA run was 0.01.³⁹ For the PSI-BLAST search, the inclusion threshold was set to 10^{-5} , the number of iterations is set to 10, and the final match threshold is set to 10^{-4} .⁴³ Since the threshold values used for the PSI-BLAST run are somewhat conservative,⁴³ in all the cases, the number of ORFs assigned by FASTA is greater than by PSI-BLAST. To enrich the sequence information in the PSI-BLAST iteration, PDB sequences were then combined with the Swiss-Prot, trEMBL³⁷ and genome sequences in the KEGG database (see the fourth column from the right in Table I, the same threshold values are used). Now PSI-BLAST because of its iterative approach, on average results in a 10.4% increase in the number of structure assignments. GTOP and PEDANT use PSI-BLAST, and GTOP tends to assign more structures than PEDANT (with *S. cerevisiae* being the only exception). Compared to GTOP, structure assignment by PROSPECTOR_Q is on average 23.6% more ORFs for a genome. Figure 2 shows the growth of fraction of ORFs in a genome to which a structure is assigned. There is around a 20% increase of the assignment between the year 2000 and 2002, and a similar big leap was made by PROSPECTOR_Q.

In contrast to the high ratio of ORF coverage of our structure assignment, the amino acid coverage obtained just by counting the total number of residues assigned to templates relative to the number of residues in the genome (the fifth column in Table I) to a genome still remains at the level of 30% for eukaryotic genomes and at 50% for prokaryotic genomes. Interestingly, the average coverage for an individual protein is 62.9%. This implies that additional structure modeling procedures are still needed for unaligned regions to obtain the whole structure of an ORF, which is *raison d’être* of an *ab initio* folding algorithms⁴⁴ despite the fairly successful ability of threading to identify structurally related proteins. This average range of coverage generally results in buildable models.⁴⁵

Effect of the Number of Sequences Used to Build the Profiles

Since the inputs to the PROSPECTOR_Q are MSAs, it is of interest to see how the number of homologous sequences

TABLE I. ORFs With Assigned Structures

Organism	Genome size (nt)	Total number of protein ORFs	ORFs with assigned structure (%)	Amino acid coverage (%)	Average coverage Per ORF (%)	FASTA ^a (%)	PSI-BLAST-PDB ^b (%)	PSI-BLAST-PDBgenes ^c (%)	GTOP (%)	PEDANT (%)	Gerstein's group ^b
<i>M. genitalium</i>	580074	484	387 (80.0)	48.1	66.3	231 (47.7)	205 (42.4)	269 (55.6)	273 (56.4)	259 (53.5)	214 (44.2)
<i>E. coli</i>	4639221	4289	3356 (78.2)	50.2	66.8	1660 (38.7)	1516 (35.3)	1724 (40.2)	2032 (58.5)	1954 (45.6)	1191 (27.8)
<i>B. subtilis</i>	4214814	4106	2988 (72.8)	47.2	66.9	1465 (35.7)	1314 (32.0)	1780 (43.4)	1947 (60.2)	1963 (47.7) ^f	1121 (27.3) ⁱ
<i>A. aeolicus</i>	1551335	1522	1297 (85.2)	48.0	66.2	646 (42.4)	592 (38.9)	783 (51.4)	827 (53.1) ^d	800 (52.6)	527 (34.6)
<i>S. cerevisiae</i>	12156306	6343	4610 (72.7)	30.0	48.1	1962 (30.9)	1804 (28.4)	2473 (39.0)	2694 (42.5) ^e	2766 (42.9) ^g	1699 (27.3) ^j

^aThe E-value threshold is set to 0.01.

^bThe inclusion threshold is set to 10^{-5} , the number of iteration is set to 10^{-4} , and the final match threshold is set to 10^{-4} . The PDB is scanned.

^cSwiss-Prot, trEMBL, and genome sequences in the KEGG database are combined with the PDB to enrich sequence profiles. The same thresholds are used as in footnote b.

^dThe total number of ORFs used in their analysis is 1556.

^eThe total number of ORFs is 6346. f.

^fIn total 4112 ORFs are used in their analysis.

^gIn total 6449 ORFs are used.

^hThe analysis was done in 2000.

ⁱIn total, 4100 ORFs are used.

^jIn total 6218 ORFs are used.

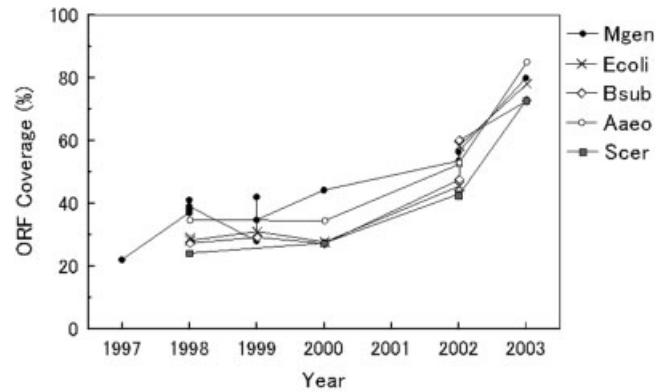


Fig. 2. Growth of the fraction of ORFs with an assigned structure in each genome. Solid circles, *M. genitalium*; crosses, *E. coli*; open diamonds, *B. subtilis*; open circles, *A. aeolicus*; gray squares, *S. cerevisiae*. Below, sources of data points used in the plot are listed for each genome. The year (i.e., X-axis in the plot) is shown in the parentheses. The data points used for *M. genitalium*: Fisher and Eisenberg (1997);²¹ Huynen et al. (1998);²² Teichmann et al. (1998);²³ Rychlewski et al. (1998);²⁴ Wolf et al. (1998);⁵⁶ Müller et al. (1999);²⁶ Jones (1999);²⁷ Salamov et al. (1999);²⁸ Gerstein et al. (2000, from the group web page, the year shows the date when the file was last modified); PEDANT (2002, the year when we retrieved the data from the database); GTOP (2002, the year when we took the data from the database); and our current result in Table I (2003). For *E. coli*: Rychlewski et al. (1998);⁵⁷ Wolf et al. (1998);⁵⁶ Salamov et al. (1999);²⁸ Gerstein et al. (2000); PEDANT (2002); GTOP (2002); our result (2003). For *B. subtilis*, *A. aeolicus*, and *S. cerevisiae*: Wolf et al. (1998);⁵⁶ Salamov et al. (1999 only for *B. subtilis*);²⁸ Gerstein et al. (2000); PEDANT (2002); GTOP (2002); our result (2003).

in the alignments affects the resulting ability to confidently assign structures. Figure 3(A) and (B) shows the distribution of the number of sequences in the close and distant sequence profiles respectively, and the corresponding fraction of structure assignments. It is obvious that the fraction of ORFs that have a structure assigned grows as the number of sequences used to construct either profile increases (solid circles); indeed, when the number of sequences in the close profile [Fig. 3(A)] exceeds 15, in more than 90% of the cases, a template structure is assigned to the ORF.

There are two reasons why this happens: the first and simpler reason is that proteins in dominant families have a larger possibility that the structure of a family member has been solved, i.e., closely homologous to the target sequence. In Figure 3(A) and (B), this case is shown in open diamonds for both close and distant profiles. The second reason is that enrichment of sequence information makes it possible to detect a structure that does not have apparent homology to the query sequence. This is illustrated by the fact that the fraction of homologous template proteins does not exceed around 70% [see Fig. 3(A)], which shows that enrichment of sequence information using the profile is a very important factor to drive the fraction of the structure assignment up to almost 100% [Fig. 3(A)] when structural information is included in the threading scoring function. As shown clearly in Figure 3(A), for 65.2% of the cases a template structure is assigned to ORFs when there is not a single homologous sequence in the close profile. Thus, only distant sequences are used in the detection. In Figure 3(B), when there is also no sequence detected in the

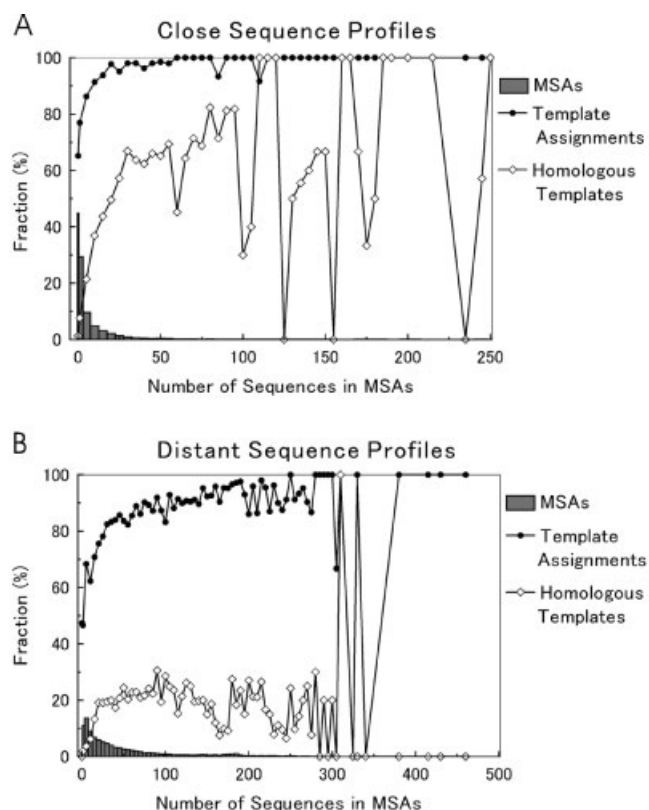


Fig. 3. The number of sequences in MSAs. **A:** The close (35–90%) profile, **B:** the distant (e -value < 10) profile. The gray bars show a histogram of the number of sequences found in a given type of MSA. The bin of the bars is set to five, except for the two most left bars, i.e., the most left bars, zero sequences; the second bar, 1–4 sequences. This is to separately show the number of cases with no homologous sequences. Solid circles, the fraction of the fold assignments to the MSAs. Open diamonds, the fraction of assigned template structures which are homologous (> 35% sequence identity) to the target ORF sequence.

e_{10} profile (i.e., only the target ORF sequence), still 47.4% of the ORFs have a template assignment.

Comparison to Other Methods

Since we have assigned substantially more template structures to each genome than other methods (Table I), our curiosity leads us to investigate the nature of newly assigned structures. In Table II, ORFs in each genome were divided into two groups, namely, the ones that have a template structure assigned only by PROSPECTOR_Q (“newly assigned ORFs”), and the ones that have a template assigned also by the other methods in Table I (i.e., FASTA, PSI-BLAST against the PDB, PDB+genes, GTOP and PEDANT) (previously assigned ORFs, termed here “preassigned”). Then, the number of distinct template proteins assigned to the two groups of ORFs is compared. The set of template proteins of the newly assigned ORFs does not overlap much with those of the previously assigned ORFs. This fact indicates that PROSPECTOR_Q tends to assign new template proteins rather than assigning the same templates to more ORFs; these newly assigned ORFs account for around 30% of the total variety of assigned templates in a genome. The average Z -score of the cases when only PROSPECTOR_Q assigned a tem-

plate is 17.31, compared to 34.74 when other methods also made a template assignment. The average number of homologous sequences in the close (distant) profile is 1.58 (27.2) and 10.29 (74.8) for the former and the latter cases, respectively. This clearly shows that the threading algorithm is making rather difficult assignments (with lower Z -score on average) thereby compensating for lack of close homologous sequence information.

In Figure 4, structure assignments from PROSPECTOR_Q, and GTOP,³² and PEDANT³¹ are compared. We have judged agreement of a pair of structure assignments by one of three ways: The first is to align two assigned structures using our recently developed protein structure alignment program, SAL⁴⁶ (<http://www.bioinformatics.buffalo.edu/resources/sal/>). A pair of fold assignments is considered to be the same if the aligned region of the two structures covers more than 80% of the smaller protein of the pair with an RMSD less than 6.5 Å [“A” in Fig. 4(A)]. The second way is to refer to the CATH database; if two assigned structures belong to a same topology defined in the CATH database, both structures are considered to be the same (“C”). The last way is to refer the SCOP⁴⁷ database in a similar fashion (“S”). Here, only those ORFs that all three procedures have assigned a structure are considered. GTOP and PEDANT consistently show the highest agreement [black histogram in Fig. 4(A)]. This may be because both procedures employ PSI-BLAST. Figure 4(B) shows the distribution of Z -scores from PROSPECTOR_Q, with ORFs classified by the three types of agreement of structure assignments. The average Z -score of the ORFs with the same assigned fold by all three methods is 37.6, that of ORFs to which a different fold was assigned only by PROSPECTOR_Q is 24.4, and that of ORFs with three differently assigned folds is 25.1. When the Z -score is less than 16, assignments by PROSPECTOR_Q do not agree with GTOP and/or PEDANT for more than half of the cases, but this disagreement decreases as the Z -score grows. When the Z -score exceeds 20.0, all three assignments agree in 92.1% of the cases. In the benchmark test, a Z -score 15.0 is the threshold above which the threading alignments are very reliable; namely 81.6% of the cases one of the top five scoring templates has an RMSD of less than 6.5 Å,²⁹ any errors if present typically involve the orientations of the N- or C-terminal secondary structural elements. One of the reasons is because the statistical significance of sequence identity between a query ORF and its template grows with the increase of the Z -score, so that the template also becomes visible by sequence-based methods.

Fold Distribution

Figure 5 shows the fold distribution of the five genomes, with the top five topologies and architectures shown in Table III. The fold classification is taken from the CATH database. A predicted structure of an ORF is considered to have the fold of the protein structure when the aligned part between the query ORF and the structure covers more than 60% of the structure. Together with the fact that not all the proteins in the PDB are included in CATH database, 89.1, 92.0, 91.5, 91.9, 90.0% of the PROSPECTOR_Q

TABLE II. New Templates Assigned by PROSPECTOR_Q[†]

	(A + B + C) Number of different templates assigned by PROSPECTOR_Q	(B) Number of different templates assigned both to newly assigned ORFs and preassigned ORFs (%) ^a	(A) Number of different templates assigned only to newly assigned ORFs (%) ^a
<i>M. genitalium</i>	323	15 (4.6)	78 (24.1)
<i>E. coli</i>	1527	336 (22.0)	506 (33.1)
<i>B. subtilis</i>	1309	258 (19.7)	438 (33.5)
<i>A. aeolicus</i>	849	104 (12.2)	231 (27.2)
<i>S. cerevisiae</i>	1855	426 (23.0)	685 (36.9)
All organisms	2692	1140 (42.3)	797 (29.6)

[†]ORFs in each genome were divided into two groups, “newly assigned ORFs”: the ones that have a template structure assigned only by PROSPECTOR_Q, and “preassigned ORFs”: the ones that have a template assigned also by the other methods in Table I. Then the number of different templates that were assigned (A) only to newly assigned ORFs; (B) both to newly assigned ORFs and preassigned ORFs; (C) only to preassigned ORFs; are counted (i.e., a Venn diagram). (A) + (B) + (C) gives the total number of different templates assigned to a genome. a) Percent of the total number of different templates assigned to the genome (i.e., (A + B + C) is the denominator).

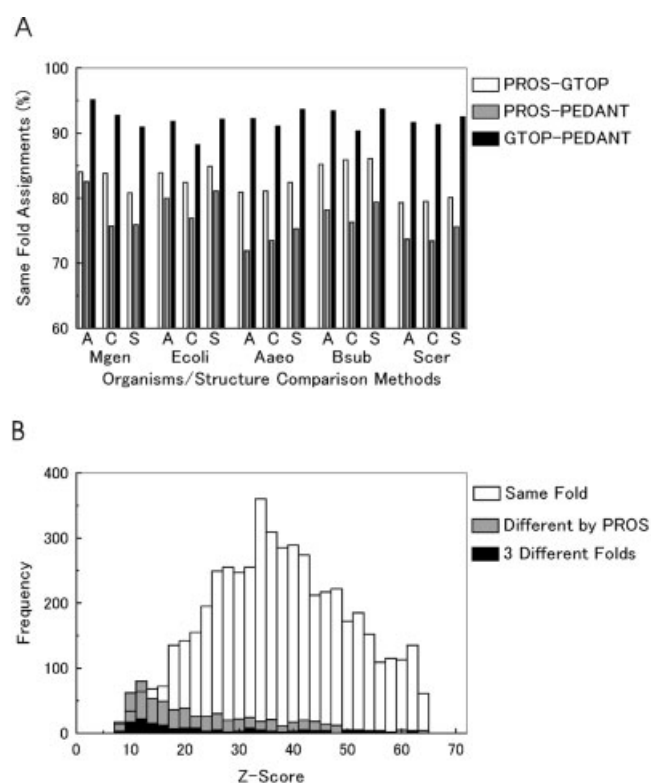


Fig. 4. Comparison of the structure assignment by PROSPECTOR_Q, GTOP and PEDANT. **A:** For each five organisms, the fraction of the agreement of structure assignment between two methods is shown: white histogram, comparison between PROSPECTOR_Q and GTOP; gray histogram, comparison between PROSPECTOR_Q and PEDANT; black histogram, comparison between GTOP and PEDANT. The comparison of assigned structure was carried out in three ways: A, our structure alignment program SAL⁴⁶ was used to superimpose two assigned structures; C, CATH database; S, SCOP database is. See text for additional details. **B:** Z-score distribution of the assigned structures. White histogram, the Z-score distribution of the ORFs to which all the three methods assigned equivalent template structures; gray bar, that of the ORFs to which only PROSPECTOR_Q assigned a different structure from the other two methods; black bar, that of ORFs to which all three methods assigned different structures.

assigned structures of *M. genitalium*, *E. coli*, *B. subtilis*, *A. aeolicus*, and *S. cerevisiae*, respectively, are counted in these statistics. The fold distribution of the five organisms

is quite similar, with on average 53.9% of the proteins predicted to be in the α/β class. Table III(A) shows the ranking of top five abundant architectures. It is very clear that 3-layer and 2-layer $\alpha\beta$ sandwiches, and orthogonal helical bundles are the most abundant in all the genomes. Table III(B) shows the ranking of the most abundant topologies, the next hierarchy of CATH classification scheme. The most abundant topology is the Rossmann fold (89) in all the organisms, and the TIM barrel (33), $\alpha\beta$ plaits (31), immunoglobulin-like (22), Arc repressor mutant subunit A (orthogonal α helical bundle structure) (18), OB fold (12) follows.

As Gerstein has pointed out,⁴³ these abundant topologies are adopted by proteins of various functions, i.e., they are multi-functional topologies: the number of different EC numbers (the first three digits) is shown in the parentheses. It is also interesting that the kinase fold enters as the fifth abundant topology of *S. cerevisiae*. There are 101 ORFs which hit a kinase fold. Among them, 87 genes are indeed a kind of kinase, with 57 genes functioning as a serine/threonine protein kinase (according to the KEGG database). There are eight ORFs of unknown function.

Function Prediction via Assigned Fold

One of the most important possible applications of a genome scale structure prediction is inferring protein function from their predicted structure, following the “sequence–structure–function” paradigm.^{4,48} As shown in Figure 1, there are many protein topologies that have a one-to-one relationship to enzyme function (EC number); these we termed “single-enzyme function” topologies. The idea here is to investigate the possibility of transferring the function of a template protein having a single-function topology to an assigned ORF.

Analyzing the structure assignment by PROSPECTOR_Q to the five genomes, there are 3460 ORFs of known function (according to KEGG) to which “single-enzyme function” template proteins are assigned. We have counted only ORFs whose sequences were covered by more than 60% by the selected template protein. Among these, 147 ORFs (i.e., 4.2% false assignment rate) have an apparently different function according to KEGG as compared to that

TABLE IIIA. Top Five Abundant Architectures in Genomes

	<i>M. genitalium</i> (%) ^a	<i>E. coli</i>	<i>A. aeolicus</i>	<i>B. subtilis</i>	<i>S. cerevisiae</i>
1	3-Layer $\alpha\beta\alpha$ sandwich 3.40 ^b (25.9)	3-Layer $\alpha\beta\alpha$ sandwich 3.40 (27.9)	3-Layer $\alpha\beta\alpha$ sandwich 3.40 (27.1)	3-Layer $\alpha\beta\alpha$ sandwich 3.40 (28.8)	3-Layer $\alpha\beta\alpha$ sandwich 3.40 (19.7)
2	2-Layer sandwich 3.30 (15.8)	Orthogonal bundle 1.10 (12.9)	2-Layer sandwich 3.30 (15.3)	Orthogonal bundle 1.10 (14.0)	2-Layer sandwich 3.30 (14.6)
3	Orthogonal bundle 1.10 (11.7)	2-Layer sandwich 3.30 (12.0)	Orthogonal bundle 1.10 (9.9)	2-Layer sandwich 3.30 (12.5)	Orthogonal bundle 1.10 (14.3)
4	(Partially classified) 6.1 (7.1)	$\alpha\beta$ Complex 3.90 (5.9)	(Partially classified) 6.1 (6.2)	(Partially classified) 8.1 (5.4)	Up-down bundle 1.20 (6.9)
5	Up-down bundle 1.20 (6.0)	(Partially classified) 8.1 (5.3)	$\alpha\beta$ Complex 3.90 (5.9)	$\alpha\beta$ Complex 3.90 (5.3)	(Partially classified) 6.1 (6.5)

^aPercentage of the assigned CATH domains in the genome is shown in the parentheses. A domain in CATH is assigned to an ORF when the alignment of the two proteins covers larger than 60% of the length of the CATH domain. The percent of ORFs in each genome to which CATH domains are assigned is: *M. genitalium*: 66.2%; *E. coli*: 68.3%; *B. subtilis*: 65.4%; *A. aeolicus*: 66.4%; *S. cerevisiae*: 63.9%.

^bTwo-digit CATH code for the architecture is shown after the name of the architecture.

TABLE IIIB. Top Five Abundant Topologies (CATH) in Genomes

	<i>M. genitalium</i> (%)	<i>E. coli</i>	<i>A. aeolicus</i>	<i>B. subtilis</i>	<i>S. cerevisiae</i>
1	Rossmann fold 3.40.50 ^a (18.8)	Rossmann fold 3.40.50 (18.1)	Rossmann fold 3.40.50 (18.6)	Rossmann fold 3.40.50 (18.4)	Rossmann fold 3.40.50 (12.8)
2	$\alpha\beta$ plaits 3.30.70 (5.2)	$\alpha\beta$ plaits 3.30.70 (4.7)	$\alpha\beta$ plaits 3.30.70 (6.4)	$\alpha\beta$ plaits 3.30.70 (5.7)	$\alpha\beta$ plaits 3.30.70 (4.4)
3	Arc repressor mutant subunit A 1.10.10 (3.5)	TIM barrel 3.20.20 (4.1)	TIM barrel 3.20.20 (3.6)	Arc repressor mutant subunit A 1.10.10 (4.0)	Immunoglobulin-like 2.60.40 (2.8)
4	OB fold 2.40.50 (2.7)	Arc repressor mutant subunit A 1.10.10 (4.0)	Arc repressor mutant subunit A 1.10.10 (2.5)	TIM barrel 3.20.20 (3.9)	Arc repressor mutant subunit A 1.10.10 (2.1)
5	Aspartyl tRNA synthetase, subunit A, domain 2 3.40.690 (2.5)	Immunoglobulin-like 2.60.40 (3.0)	OB fold 2.40.50 (2.1)	Aminopeptidase 3.40.630 (1.9)	Kinase 3.30.200 (2.7)

^aThree digit CATH code for the topology is shown after the name of the topology. Percentage of the assigned CATH domains in the genome is shown in the parentheses.

of the PROSPECTOR_Q assigned template protein. The small false positive rate suggests that the concept of function assignment via the single-function fold works fairly well. There are 71 template proteins (17.3% of the total number of single function templates) involved in these apparently false positive assignments. We have rejected those 71 template proteins from the list of “single-enzyme function” template proteins resulting in 339 single-function template proteins, which were used to annotate function via predicted structure to ORFs for unknown function. This step constitutes the final screening of “single-enzyme function” topologies.

Next, we have checked those cases when template proteins of “single-enzyme function” topologies are assigned to ORFs of unknown function. ORFs of unknown function are selected by the keywords “unknown” or “hypothetical” in the DEFINITION field of the KEGG database. In the five genomes, a total of 120 ORFs of unknown function are assigned to template proteins with a “single-enzyme function” topology. Since the list of these template proteins is based on current databases which are still expanding and only enzyme function is considered (Fig. 1),

we are reluctant to naively transfer function through the assigned template proteins without further investigation. Therefore, we have checked if PROSITE⁴⁹ or Pfam¹⁷ patterns, or active site residues described in the literature of the template protein are shared with the ORFs and are well aligned by PROSPECTOR_Q. Among these 120 cases, there are 40 cases where we could find some supporting evidence, i.e., those aligned motifs/active site residues in their threading alignments. In 35 cases out of the 40 cases, either GTOP or PEDANT assigns equivalent template proteins. Below we show four cases where neither GTOP nor PEDANT has assigned any structure/function, but where there is some evidence that the function prediction may be correct.

yozI. A *B. subtilis* ORF, yozI hit 1b13A, rubredoxin with a Z-score of 27.66. The rubredoxins are iron-sulfur proteins, featuring a single Fe(S-Cys)₄ site in a protein.⁵⁰ As shown in Figure 6, yozI has all four cysteine residues in two loops aligned to the cysteine ligands in the rubredoxin family. YozI also has Gly10, which is said to be important for the maintenance of hydrogen-bonding interactions around Fe(S-Cys)₄ site (In 1b13A, Gly10 is artificially

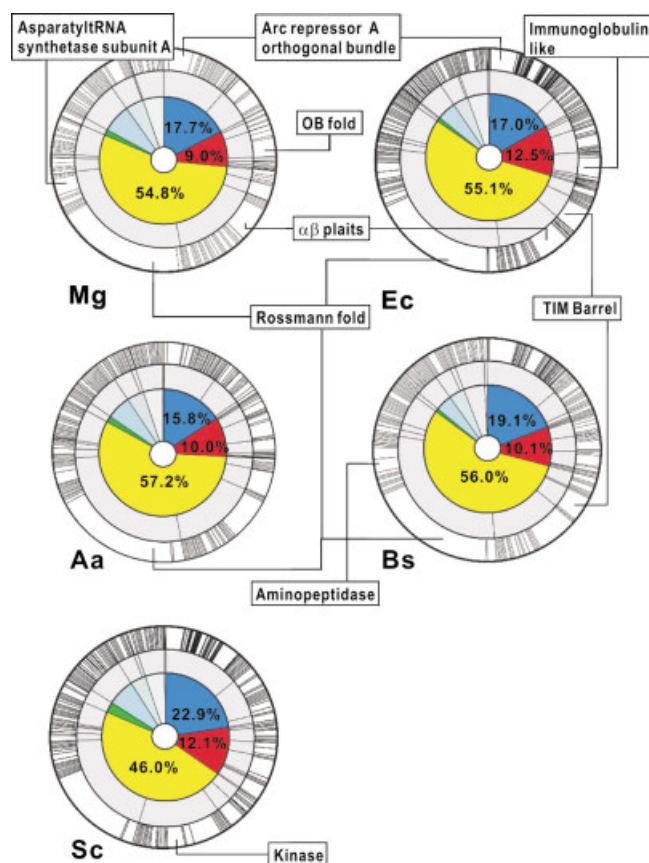
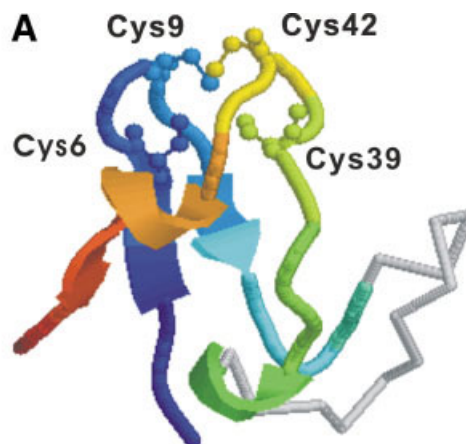


Fig. 5. Fold distribution of the five organisms (the CATH wheel⁵⁸). This representation comprises three concentric pie charts. The most inner circle in color shows the class [C] of template proteins hit for a genome: blue, mainly α ; red, mainly β ; yellow, mixed $\alpha\beta$; green, low secondary structure content. The rest of them, colored in pale blue shades, are classified to the class number 5–9, which are the preliminary classifications in CATH. The middle circle represents the architecture [C.A], and the outer circle represents the topology [C.A.T]. The angle defined by any segment is proportional to the number of assigned template proteins in the category. Names of the top five dominant topologies are shown. Abbreviations of the five organisms are: Mg, *M. genitalium*; Ec, *E. coli*; Aa, *A. aeolicus*; Bs, *B. subtilis*; Sc, *S. cerevisiae*.

mutated to Ala). 1b13A has a PROSITE pattern PS00202 (rubredoxin signature), [LIVM]-x(3)-W-x-C-P-x-C-[AGD], in the region of residues 33–43. YozI only lacks the first residue and the tryptophan from the PROSITE pattern, but has the rest of the residues specified by the PROSITE pattern.

YrdN, a protein in *B. subtilis* hits 1bjpA, 4-oxalocrotonate tautomerase with a Z-score of 18.22. Three functionally important residues in the catalytic site, Pro1, Arg39 and Arg61⁵¹ are aligned between 1bjpA and yrdN. yrdN is also recognized by the Pfam¹⁷ signature as a tautomerase.

b3555 (yiaG), an *E. coli* ORF hits 1b0nA, SinR transcription regulator with a Z-score of 19.90. Since the lengths of both proteins are similar (1b0nA, 103 residues; b3555, 96 residues), and both hit Pfam HTH_3 (a helix-turn-helix type transcription factor). Babu & Teichmann list this protein as an unknown lambda repressor-like DNA-binding domain.⁵²



B

```

RUBR_CLOST MTKYVCTVGGVYDPEVGD PDNNINPGTISFDI PRDWWCPGQGVGKDQFESEA-
RUBR_CLOPE MKKFDICDVGVIYDPAVGD PDNGVEPGTEFKDI PDDWVCPGQGVDRSQFSETE-
RUBR_HELMO MKKYVCTVGGVYDPAKGD PDHGIAPGTAFADEL PADWVCPGQGVSKDFEEL--
RUBR_BUTME MOKYVCTVGGVYDPAVGD PDNGVAPGTAFADEL PADWVCPGQGVSKDFEPEA-
RUBR_CHLLT MOKYVCTVGGVYDPAKGD PDHGIAPGTAFADEL PADWVCPGQGVSKDFEPEA-
RUB3_CHLTE MOKYVCTVGGVYDPAKGD PDHGIAPGTAFADEL PADWVCPGQGVSKDFEPEA-
RUBR_DESGI MDIYVCTVGGVYDPAKGD PDHGIAPGTAFADEL PADWVCPGQGVSKDFEPEA-
RUBR_DESMV MKKYVCTVGGVYDPAKGD PDHGIAPGTAFADEL PADWVCPGQGVSKDFEPEA-
RUBR_PYRFU --AKWVCKIIGVYIDDEAGD PDNGISPGTKFEDLPDDWVCPGQGVSKDFEPEA-
RUBR_CLOTS MEKWQCTVGGVYDPEVGD PPTQNIIPPGTKFEDLPDDWVCPGQGVSKDFEPEA-
RUB2_CHLTE MEQWVCTVGGVYINPFGTDPGDPGDIIPAGTISFESL PDDWVCPGQGVSKDFEPEA-
RUBR_METTH MKKYVCTVGGVYDPEKGEPRDTIPPGTIPFEDLPETWVCPGQGVSKDFEPEA-
RUBR_ACICA MKKYVCTVGGVYDPAEAGWPDGIAFGTKWEDI PDDWVCPGQGVSKDFEPEA-
1b13A MKKYVCTVGGVYINPFGTDPGDPGDIIPAGTISFESL PDDWVCPGQGVSKDFEPEA-
yozI (prospector) MKLYTCTVGGVYINPFGTDPGDPGDIIPAGTISFESL PDDWVCPGQGVSKDFEPEA-
yozI (align0) MPK... ..GLIQNPLNNDGEYQ...FD... ..DFEYGYSEDHVV

```

Fig. 6. Threading alignment of yozI (*B. subtilis*) to a template, 1b13A (rubredoxin). **A**: The structure of 1b13A. Aligned parts to yozI are shown in color (blue, N-terminus; red, C-terminus). Four cysteine residues that constitute Fe(S-Cys)₄ site are shared with yozI, and shown in ball & sticks model. **B**: A multiple alignment of rubredoxin sequences. The sequences are taken from Swiss-Prot. These sequences share a sequence identity of 35% or more to 1b13A, and less than 90% to each other. Two alignments of yozI are shown at the bottom: One aligned by PROSPECTOR_Q, and the other aligned by align0⁵⁹ to 1b13A. The secondary structure, i.e. helices and strands shown in gray boxes and white boxes, respectively. For 1b13A, DSSP⁶⁰ is used for the definition of the secondary structure. For yozI, Pspred⁶¹ is used to predict the secondary structure. The four cysteine residues are marked in rectangles.

b2399 (yfeD), an *E. coli* ORF hits 1adr, p22 c2 repressor with a Z-score of 10.4. It also hits two Pfam signatures, bacterial regulatory protein lacI and HTH_3. Babu and Teichmann annotate this protein as unknown lambda repressor-like DNA-binding domain.⁵²

DISCUSSION

In this paper, we have reported the genome-scale structure prediction of five organisms using PROSPECTOR_Q. The new features of PROSPECTOR_Q are that it incorporates a better treatment of gaps at the both termini of the alignment and that following Karplus,³⁵ the score is evaluated as the energy difference between the best score of the target sequence aligned to the template and the reversed sequence aligned to the template. Using the Z-score threshold established in an extensive benchmark test, we could assign structures to 72–85% of the ORFs in the examined genomes. Compared to earlier studies by several authors, this fraction is greater by almost 20% on average. About half of this improvement comes from the assignment of ORFs to new template proteins that were not assigned before (Table II), and many of these tem-

plates are not closely homologous [Fig. 3(A, B)]. Thus, the PROSPECTOR_Q threading algorithm is capturing structural features of the template structures. Actually, even when no homologous sequences are found for distant profiles [Fig. 3(B)], PROSPECTOR_Q finds a template above the Z-score threshold for 47.4% of the cases.

On the other hand, comparing the assigned structures by PROSPECTOR_Q, our assignment matches around 80% of the GTOP and PEDANT fold assignments [Fig. 4(A)]; the mismatch ratio grows as the Z-score deteriorates [Fig. 4(B)]. At some point, this mismatch of structure predictions is inevitable because of the different characteristics/limitations of each method. Another problem is the accuracy of threading alignments; when the Z-score is not very high, an accurate alignment may not be obtained even if the fold itself is correctly identified. Therefore, although our algorithm was tested on the benchmark set, where the fold assignment is 95% accurate, for practical use of a threading-based protein model, one should still carefully check the validity of the model.

Although the fraction of ORFs with an assigned template structure is large, in the majority of the cases, the predicted structure does not cover the entire region of the ORF sequence, with on average around 60% is aligned (Table I). To deal with the problem of gaps in predicted structure, we have developed procedures to build a model of an entire molecule of a protein based on threading alignments⁵³ containing gaps.⁴⁵ The resulting predicted structures can be further utilized in protein-ligand docking⁵⁴ or protein-protein interaction⁹ prediction.

Hegyí and Gerstein⁵⁵ have done a survey on the relationship between protein fold and function. Through a similar analysis, we have identified CATH topologies, which are associated with only one function (more precisely, three digits of an EC number), which we termed here “single-enzyme function” topologies. Then, as a further step, we have tried to transfer function when an ORF has a single-function template protein assigned. This simple procedure worked fairly well for known cases, and we have made some predictions based on this prediction scheme. The idea of transferring function through predicted structure is based on the observation that protein structure is more conserved than sequence, so that evolutionary history might be better tracked by looking at the structure rather than its sequence.³³ But practically, it should be kept in mind that the current list of “single-enzyme-function” topologies may include false positives, since this list is based on current versions of databases and also non-enzyme functions are not counted. Therefore it is necessary that a function assignment be supported by other sources, such as experimental evidence or conservation of functional residues. If used properly, we believe that due to its simplicity, it might be useful for the preliminary screening process associated with massive genome-scale functional annotation.

Based on the present results and due to the fact that an increasing number of template protein structures will be available from the expansion of PDB, there is no doubt that threading algorithms will continue to play an ex-

tremely important role in the progress of structural genomics.

REFERENCES

1. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res* 2002;30:42–46.
2. Burley SK. An overview of structural genomics. *Nat Struct Biol* 2000;7 Suppl:932–934.
3. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. *Nat Biotechnol* 2000;18:283–287.
4. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the *Escherichia coli* genome using the sequence- to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J Mol Biol* 1998;282:703–711.
5. Zhang L, Godzik A, Skolnick J, Fetrow JS. Functional analysis of the *Escherichia coli* genome for members of the alpha/beta hydrolase family. *Fold Des* 1998;3:535–548.
6. Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J. Ab initio protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. *Proc Natl Acad Sci USA* 2002;99:5993–5998.
7. Sanchez R, Sali A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc Natl Acad Sci USA* 1998;95:13597–13602.
8. Yamaguchi A, Iwadate M, Suzuki E, Yura K, Kawakita S, Umeyama H, Go M. Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species. *Nucleic Acids Res* 2003;31:463–468.
9. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* 2002;49:350–364.
10. Lu L, Arakaki AK, Lu H, Skolnick J. Multimeric Threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res* 2003;13:1146–1154.
11. Smith GR, Sternberg MJ. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol* 2002;12:28–35.
12. Vajda S, Vakser IA, Sternberg MJ, Janin J. Modeling of protein interactions in genomes. *Proteins* 2002;47:444–446.
13. Jones DT. Protein structure prediction in the postgenomic era. *Curr Opin Struct Biol* 2000;10:371–379.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
15. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* 2000;9:232–241.
16. Eddy SR. Hidden Markov models. *Curr Opin Struct Biol* 1996;6:361–365.
17. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res* 2002;30:276–280.
18. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 1998;14:846–856.
19. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins* 2001;42:319–331.
20. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D. TOUCHSTONE: a unified approach to protein structure prediction. *Proteins* 2003;53:469–479.
21. Fischer D, Eisenberg D. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc Natl Acad Sci USA* 1997;94:11929–11934.
22. Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P. Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J Mol Biol* 1998;280:323–326.
23. Teichmann SA, Park J, Chothia C. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci USA* 1998;95:14658–14663.
24. Rychlewski L, Zhang B, Godzik A. Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold Des* 1998;3:229–238.

25. Wolf YI, Brenner, Steven, E., Bash, Paul A., & Koonin, Eugene V. Distribution of protein folds in the three superkingdoms of life. *genome research* 1999;9:17–26.
26. Muller A, MacCallum RM, Sternberg MJ. Benchmarking PSI-BLAST in genome annotation. *J Mol Biol* 1999;293:1257–1271.
27. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 1999;287:797–815.
28. Salamov AA, Suwa M, Orengo CA, Swindells MB. Genome analysis: Assigning protein coding regions to three-dimensional structures. *Protein Sci* 1999;8:771–777.
29. Skolnick J, Kihara D. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Forthcoming* 2003.
30. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;85:2444–2448.
31. Frishman D, Mokrejs M, Kosykh D, Kastenmuller G, Kolesov G, Zubrzycki I, Gruber C, Geier B, Kaps A, Albermann K, Volz A, Wagner C, Fellenberg M, Heumann K, Mewes HW. The PEDANT genome database. *Nucleic Acids Res* 2003;31:207–211.
32. Kawabata T, Fukuchi S, Homma K, Ota M, Araki J, Ito T, Ichiyoshi N, Nishikawa K. GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res* 2002;30:294–298.
33. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. *Structure* 1997;5:1093–1108.
34. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins* 2000;38:3–16.
35. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry. *Proteins* 2003;51:504–514.
36. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
37. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;31:365–370.
38. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;85:2444–2448.
39. Pearson WR. Empirical statistical estimates for sequence similarity searches. *J Mol Biol* 1998;276:71–84.
40. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680.
41. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;28:304–305.
42. Bukhman YV, Skolnick J. BioMolQuest: integrated database-based retrieval of protein structural and functional information. *Bioinformatics* 2001;17:468–478.
43. Hegyi H, Lin J, Greenbaum D, Gerstein M. Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds. *Proteins* 2002;47:126–141.
44. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci USA* 2001;98:10125–10130.
45. Zhang Y, Skolnick J. Assembly of protein tertiary structures from substructures of weakly scoring threading templates. *Forthcoming* 2003.
46. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. *J Mol Biol* 2003;334:793–802.
47. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 2002;30:264–267.
48. Norin M, Sundstrom M. Structural proteomics: developments in structure-to-function predictions. *Trends Biotechnol* 2002;20:79–84.
49. Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. *Nucleic Acids Res* 1999;27:215–219.
50. Maher MJ, Xiao Z, Wilce MC, Guss JM, Wedd AG. Rubredoxin from *Clostridium pasteurianum*. Structures of G10A, G43A and G10VG43A mutant proteins. Mutation of conserved glycine 10 to valine causes the 9-10 peptide link to invert. *Acta Crystallogr D Biol Crystallogr* 1999;55:962–968.
51. Taylor AB, Czerwinski RM, Johnson WH, Jr., Whitman CP, Hackert ML. Crystal structure of 4-oxalocrotonate tautomerase inactivated by 2-oxo-3-pentynoate at 2.4 Å resolution: analysis and implications for the mechanism of inactivation and catalysis. *Biochemistry* 1998;37:14692–14700.
52. Madan Babu M, Teichmann SA. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res* 2003;31:1234–1244.
53. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 2001;44:133–149.
54. Wojciechowski M, Skolnick J. Docking of small ligands to low-resolution and theoretically predicted receptor structures. *J Comput Chem* 2002;23:189–197.
55. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* 1999;288:147–164.
56. Wolf YI, Brenner SE, Bash PA, Koonin EV. Distribution of protein folds in the three superkingdoms of life. *Genome Res* 1999;9:17–26.
57. Rychlewski L, Zhang B, Godzik A. Functional insights from structural predictions: analysis of the *Escherichia coli* genome. *Protein Sci* 1999;8:614–624.
58. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM. Protein folds and functions. *Structure* 1998;6:875–884.
59. Myers EW, Miller W. Optimal alignments in linear space. *Comput Appl Biosci* 1988;4:11–17.
60. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
61. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.