PROTEINS: Structure, Function, and Genetics 42:23–37 (2001)

Fold Recognition and Accurate Query-Template Alignment by a Combination of PSI-BLAST and Threading

Yibing Shan, Guoli Wang, and Huan-Xiang Zhou*
Department of Physics, Drexel University, Philadelphia, Pennsylvania

ABSTRACT **A homology-based structure prediction method ideally gives both a correct fold assignment and an accurate query-template alignment. In this article we show that the combination of two existing methods, PSI-BLAST and threading, leads to significant enhancement in the success rate of fold recognition. The combined approach, termed COBLATH, also yields much higher alignment accuracy than found in previous studies. It consists of two-way searches both by PSI-BLAST and by threading. In the PSI-BLAST portion, a query is used to search for hits in a library of potential templates and, conversely, each potential template is used to search for hits in a library of queries. In the threading portion, the scoring function is the sum of a sequence profile and a 6×6 substitution matrix between predicted query and known template secondary structure and solvent exposure. "Two-way" in threading means that the query's sequence profile is used to match the sequences of all potential templates and the sequence profiles of all potential templates are used to match the query's sequence. When tested on a set of 533 nonhomologous proteins, COBLATH was able to assign folds for 390 (73%). Among these 390 queries, 265 (68%) had root-mean-square deviations (RMSDs) of less than 8 Å between predicted and actual structures. Such high success rate and accuracy make COBLATH an ideal tool for structural genomics. Proteins 2001;42:23–37.**

Key words: structure prediction; structural genomics; COBLATH; screening; secondary structure prediction

INTRODUCTION

The past few years have seen tremendous progress in assigning the fold of a protein by threading its sequence to a library of potential templates.[1–28] The recent development of PSI-BLAST[29] has significantly enhanced our ability to detect remote homologues.[30–33] Both threading and PSI-BLAST have their strengths and weaknesses. For queries that have only remote homologues as templates, it may well happen that one method will work in some instances whereas the other method will work in other instances. In this article we present a combination of the two methods. We will demonstrate the superiority of the combined approach, termed COBLATH, by the significant enhancement in the success rate of fold recognition and the higher alignment accuracy than found in previous studies.

In threading one compares a query with a library of potential templates by using a scoring function. The scoring function typically involves the sequences as well as other local parameters such as secondary structure and solvent exposure. Its advantage lies in the fact that one only has to discriminate one "true" template against a finite number (e.g., a few hundred) of other decoys. In other words, regardless of how good the query-template match is, as long as the decoys score worse than the true template, the fold recognition is successful.

The strategy of PSI-BLAST is different. The core of the method is multiple sequence alignment. The sequences of the query and the template are aligned through a series of intermediaries. In a snowball fashion, more and more remotely homologous sequences are aligned with the query. When one of these sequences belongs to a protein with a known structure, fold recognition has succeeded. In this approach, one would use a database of sequences as exhaustive as possible.

In COBLATH, we exploit the complementarity of the existing methods. Both PSI-BLAST and threading are used for fold recognition. The scoring function in our threading is the sum of the sequence profile and a 6×6 substitution matrix between predicted query and known template secondary structure and solvent exposure. The sequence profile is directly from the PSI-BLAST portion. For predicting the query secondary structure and solvent exposure we use the query sequence profile as input to a neural network. Regardless of whether the template is assigned by PSI-BLAST or threading, the query-template alignment is obtained by a new round of threading.

The utility of COBLATH was demonstrated on the structural annotation of the *Mycoplasma genitanium* (MG) and *Saccharomyces cerevisiae* (SC) Genomes. Overall, COBLATH was able to assign structural templates for 298 (62%) of the MG 479 open reading frames (ORFs) and 2883 (45%) of the 6337 SC ORFs.

Grant sponsor: NIH; Grant number: GM58187.

*Correspondence to: Huan-Xiang Zhou, Department of Physics, Drexel University, Philadelphia, PA 19104.
E-mail: hxzhou@einstein.drexel.edu

Received 1 May 2000; Accepted 24 August 2000

METHODS

PSI-BLAST Parameters and Protocol

PSI-BLAST was carried out with both e and h set to 0.5×10^{-3}. For each query, PSI-BLAST was run up to 20 rounds. The problem of drifting (i.e., sequences aligned in the first round were no longer present in a later round) was dealt with by reverting to the last round before drifting for checking output. The database consists of 348,901 sequences in Swissprot and 6,680 sequences of proteins in the Protein Data Bank with less than 95% identities (PDB95). The substitution matrix was BLOSUM62.

PSI-BLAST was run in two ways. That is, a query was used to search for hits in a library of potential templates (called qblast) and, conversely, each potential template was used to search for hits in a library of queries (called tblast). The library of potential templates in qblast consists of all the PDB95 proteins. To make tblast most effective, the 1907 nonhomologous proteins in the FSSP library[34] were used as input sequences and their structural alignments were used as part of the input.

If several templates were identified for a query by qblast, the one with the lowest h value was retained. Similarly, if several FSSP proteins were matched with the same query by tblast, the one with the lowest h value was retained. When both qblast and tblast yielded a template, we then checked whether they were structural neighbors. If they were, the template with a lower h value was selected as the true template. Otherwise no template was selected for that query.

Scoring Function of Threading

Our threading for fold recognition was carried out by the global-local dynamic programming.[35,36] That is, gap penalty was imposed for unaligned N or C terminal residues of the template but not for those of the query. The scoring function was the sum of the sequence profile and a 6×6 substitution matrix $\{S_{\mu\nu}\}$ for comparing predicted query and known template secondary structure and solvent exposure. The threading was also run in two-ways (called qthread and tthread). In qthread the query sequence profile from qblast was used to match the sequences in the fold library and in tthread the template sequence profiles from tblast were used to match the query sequence. The secondary structure and solvent exposure of each residue were represented by a variable (μ or ν) with six states: buried helix, exposed helix, buried strand, exposed strand, buried coil, and exposed coil. The elements of the 6×6 substitution matrix were

$$\{S_{\mu\nu}\} = \begin{bmatrix} 1.3 & 0.8 & -4 & -5 & -1 & -1 \\ 0.1 & 1.9 & -5 & -5 & -1 & 0 \\ -3 & -4 & 1.5 & 0.3 & 0 & -1 \\ -3 & -2 & 0.6 & 1.8 & 0 & 0.5 \\ -1 & -1 & 0 & -1 & 1.0 & 0.6 \\ -2 & 0 & -1 & 0 & 0.4 & 1.4 \end{bmatrix} \quad (1)$$

The gap penalty was 10 for gap opening and 0.6 for gap extension.

The construction of $\{S_{\mu\nu}\}$ was based on the structural alignments from FSSP and the predicted secondary structure and solvent exposure of FSSP proteins. Suppose that an FSSP protein at residue position i is structurally aligned with N_i other proteins. The predicted state μ_i of the FSSP protein at residue position i was compared to the known states ν_n of the N_i structural neighbors (at the equivalent positions). At the end, a matrix with elements $p_{\mu\nu}$ was built, where $p_{\mu\nu}$ represents the total number of times a predicted μ state is paired with a known ν state. These elements were transformed into $S_{\mu\nu}$ via $S_{\mu\nu} = \ln_2(p_{\mu\nu}\Sigma_{\mu,\,\nu}p_{\mu\nu}/\Sigma_\mu p_{\mu\nu}\Sigma_\nu p_{\mu\nu})$. This treatment of the matching of predicted and known states is nearly identical to that used for sequence alignments in obtaining the BLOSUM matrix.[37] Similar procedures have been implemented previously to calculate substitution matrices involving predicted secondary structure.[7,8] The $\{S_{\mu\nu}\}$ matrix is very stable, quickly reaching the values listed above after using less than 100 representative FSSP proteins.

Prediction of Secondary Structure and Solvent Exposure

The prediction of secondary structure and solvent exposure was made by using the query sequence profile from PSI-BLAST as input to a neural network, an idea proposed by Jones.[38] When trained on 1374 FSSP proteins and tested on the remaining 533 FSSP proteins, the overall three-state accuracy of the secondary structure prediction was 72.8%. The overall two-state accuracy of the solvent exposure prediction was 79.3%.

The accuracy of the secondary structure prediction was further improved by a second neural network. The second neural network was trained on a particular class of proteins (α, β, or mixed). If, for example, the first neural network predicts an α protein, then a second neural network trained on α proteins was used to make a final prediction. Overall the second neural network improves the three-state accuracy to 74.6%.

Screening by Secondary Structure

Our fold library for threading consists of the 1907 FSSP proteins. The threading was made efficient by recognizing the simple fact that a tertiary structural match dictates secondary structural match. By using the predicted secondary structure of a query, we reduced the number of potential templates from 1907 to a screened 100.

The screening was carried out by global-local dynamic programming. Each segment of secondary structure was represented both by a letter λ (= H, B, or C) and a number l (the length of the segment). The substitution matrix was

$$\min(l_i, l_j)\begin{bmatrix} 1.0 & -1.0 & -0.3 \\ -1.0 & 1.0 & -0.3 \\ -0.3 & -0.3 & 0.5 \end{bmatrix} - |l_i - l_j|\begin{bmatrix} 0.1 & 0.5 & 0.3 \\ 0.5 & 0.3 & 0.3 \\ 0.5 & 0.5 & 0.1 \end{bmatrix}$$

For example, when a 15-residue H segment is matched a 10-residue B segment, the score would be $10\times(-1.0)-(15-10)\times0.5 = -1.25$. The matrix in the second term was

introduced to account for gap penalty. The numerical values listed above were picked without serious optimization.

In the screening the secondary-structure segments of a protein are the basic units, hence matching of the query and the 1907 potential templates can be done much more efficiently than regular threading (which entails dealing with individual residues). In an ideal situation, all the homologues of the query would be among the screened 100.

Template Selection in Threading

In PSI-BLAST a query sequence may not have any matches with proteins in the PDB. If it does, the matches are deemed true templates (if the protocol is properly chosen). Fold recognition by threading is different. One ranks all the potential templates (in our case, the screened 100) by the scoring function and thus there is always one match that ranks the top. The question is then whether the top match should be chosen as the template.

We devised the following scheme for deciding whether a template should be selected for a particular query. It involves the per-residue score f and the Z value of the top match. The former is defined as $f = F_1/L$, where F_1 is the score of the top match and L is the sequence length of the query. The Z value is given by $(F_1 - \langle F \rangle)/\sigma$, where $\langle F \rangle$ and σ are the average and the standard deviation, respectively, of the scores of those matches that rank from the second place on down but have positive scores. Importantly, we also took into account the number, M, of structural neighbors of the top match that rank at the second to fifth places. Obviously, if structural neighbors are also ranked among the best, then the probability that the top match gets there by chance is reduced. The top match was selected to be a true template if $f > f^*$ and $Z > Z^*_M$, where the subscript signifies explicit M dependence. For qthread $f^* = 0.4$ and $Z^*_M = 5.0, 4.0, 4.0, 4.0$, and 2.0, respectively, at $M = 0$ to 4. For tthread $f^* = 0.5$ and $Z^*_M = 5.0, 5.0, 4.0, 4.0$, and 2.5, respectively. When both qthread and tthread yielded a template, the outcome of qthread was chosen as the true template if the outcome of tthread is a structural neighbor.

Query-Template Alignment

In threading for fold recognition, the focus of the search is on the library of folds and the goal is to discriminate the true template from decoys. On the other hand, in query-template alignment the focus of the search is on the different combinations of aligned fragments and gaps and the problem is to compare one combination against another. A protocol effective for fold recognition may not be good for query-template alignment. In particular, we suggest that the global-local algorithm useful for fold recognition may be bad for alignment since it forces all parts of the template sequence to be aligned with the query sequence. Consequently we obtained all the query-template alignments by the local-local algorithm. Alignment accuracy may also be improved by focusing on regions of the query and template sequences that show higher similarity (as measured by the scoring function). This can be achieved by reducing the gap penalty (thus relegating the less similar regions to gaps). The final alignment results were obtained by using a penalty of 5 for gap opening and 0.3 for gap extension (half of those used for fold recognition).

A Test Set of 533 FSSP Proteins

For the purpose of evaluating COBLATH, we selected 533 FSSP proteins as queries. Each of these proteins has at least 60% of its sequence covered by the structural alignment with a structural neighbor. The 60% coverage is perhaps a threshold for fold recognition by threading.[39] It by no means guarantees a successful fold recognition, as factors such as long gaps (either in query or in template) and low sequence identity can easily allow the template to escape detection.

An important component of COBLATH is that structural alignments from FSSP were used as input for tblast. Many of the 1907 FSSP proteins have structural alignments with one or more of the 533 test proteins. For a fair evaluation of our method, we eliminated all such alignments in running tblast.

Fold recognition was labeled successful whenever the identified template is a structural neighbor of the query. All close homologues of the query were excluded from consideration. A close homologue is a protein that shows up in the query's structural alignments but is not one of the 1907 FSSP proteins.

Other Test Sets

Two other sets of proteins appeared in previous work were studied to illustrate the generality of COBLATH and to compare with alternative methods. The first consists of 68 proteins compiled by Fischer and Eisenberg.[36] The accompanying fold library consists of 301 proteins. The second set consists of 12 proteins compiled by Kolinski et al.[40] The focus for the latter set of proteins is on the accuracy of query-template alignment.

Structural Annotation of the MG and SC Genomes

COBLATH was applied to the structural annotation of the MG and SC Genomes with minor modifications. In general stricter criteria were used in searching hits for ORF sequences. In particular, fortuitous hits due to low complexities were avoided by filtering. The Swissprot sequences were filtered with a trigger window length of 12 and trigger and extension complexities of 1.8 and 2.0. ORF sequences were filtered with a trigger window length of 12 and trigger and extension complexities of 2.2 and 2.5.

Tblast was carried out in two stages. A preliminary session of PSI-BLAST was carried out using the 1907 FSSP proteins as input sequences and their structural alignments as seeding. In the last round (up to 20) all alignments with PDB95 and Swissprot sequences were saved. These alignments were used as seeding in a second session of PSI-BLAST, where the input sequences were again the FSSP proteins but the database now consisted of genomic sequences. However, if for a particular input FSSP sequence, drift occurred in the first session of

PSI-BLAST, then alignment information from this session was discarded. Instead the structural alignment of the FSSP protein was directly used as seeding in the second session of PSI-BLAST, where the database now consisted of PDB95 and genomic sequences. In any event, if drift occurred in the second session of PSI-BLAST after the fifth round, only hits up to the third round were retained.

Very often several templates were assigned to the same ORF. These templates were ordered from low to high h values. If the ORF region (O_1) covered by the template with the lowest h value overlapped with that (O_2) covered by another template and the overlap was at least 30% of O_1 and 50% of O_2, then only the first template was retained. In that case we further checked whether the second template was a structural neighbor of the first according to the FSSP library. If so we went on to compare the rest of the templates with those already retained; otherwise the process of finding additional templates for an ORF was stopped.

When a structural template was assigned to an ORF by one component of COBLATH, there is no need to try the other three components. Hence the four components of COBLATH were applied to the SC genome in the following order: tblast, qblast, qthread, and tthread. However, to further test the reliability of COBLATH, all the four components were applied to each of the 479 MG ORF sequences.

RESULTS

Screening of Fold Library

The sequence-profile based neural network is arguably the best method for secondary structure prediction.[38] We further improved this method by an additional network that was trained on a single class of proteins. The increased accuracy has direct impact on the screening of the fold library. The performance of the screening can be simply evaluated by checking whether the query itself is among the screened 100. Of the 533 queries, all but 49 were retained by the screening after using the second neural network. In comparison, 68 would not have made into the screened 100 if only one neural network was used. We note that, when a query itself was among the screened 100, it was simply discarded.

Success Rate of Fold Recognition

The number of queries for which templates were identified was 307, 292, 330, and 307 by qblast, tblast, qthread, and tthread, respectively. Significantly, each method identified templates for a large number of queries that did not have templates assigned by any other methods. Specifically, 46 of the queries were assigned templates by qblast but not by tblast, and 31 queries were assigned templates by qblast but not by qblast. By merging the results of qblast and tblast, the total number of queries assigned templates by two-way PSI-BLAST was 336 (two queries, 1lxa and 1nfn, had their templates rejected because of conflict between qblast and tblast). Similarly, 43 of the queries were assigned templates by qthread but not by tthread, and 20 queries were assigned templates by qthread but not by qthread. By merging the results of qthread and tthread, the total number of queries assigned templates by two-way threading was 350. Of the 197 queries not assigned templates by two-way PSI-BLAST, 54 were assigned templates by two-way threading.

Overall, COBLATH identified templates for 390 of the 533 queries. The success rate thus stands at 73%. This is higher by over 10 percentage points than that of the best single method, qthread. The query-template pairs are listed in Table I.

Among the 390 templates, two (for queries 1poa and 1vmoA) identified by two-way PSI-BLAST were incorrect. Two-way threading had two false positives, incorrectly identifying 1av1A as the template for 1nfn and 1bax as the template for 1t1dA. The overall error rate of COBLATH is thus 1%.

The 143 queries for which templates were not identified were 1a2zA, 1a34A, 1a4mA, 1a62, 1a6f, 1a9v, 1abv, 1ag4, 1agqA, 1agrE, 1al3, 1amk, 1amx, 1aohA, 1ap8, 1arb, 1auz, 1avoB, 1aw8B, 1awcA, 1ax8, 1ayoA, 1azsA, 1b10, 1b3tA, 1b5tA, 1b66A, 1b77A, 1b79A, 1b8bA, 1bkb, 1bl0A, 1bmfG, 1bndA, 1bnkA, 1bteA, 1buoA, 1bv1, 1bvq, 1bw4, 1bxe, 1bxm, 1cem, 1chd, 1ct5A, 1dekA, 1dfx, 1dkgA, 1dpgA, 1dptA, 1e2aA, 1fbaA, 1fkj, 1g31A, 1gpr, 1havA, 1hce, 1hcnB, 1hiwA, 1hjrA, 1hmt, 1hulA, 1huuA, 1iibA, 1jhgA, 1jli, 1jmcA, 1jotA, 1l92, 1lfb, 1lki, 1lktA, 1mai, 1maz, 1mkaA, 1mroA, 1mroB, 1mspA, 1mtyG, 1mugA, 1ndoB, 1onrA, 1opy, 1otgA, 1ounA, 1pauA, 1pbv, 1pda, 1pdo, 1pgs, 1phm, 1prtF, 1pud, 1qfhA, 1rcb, 1regX, 1rgeA, 1rhoA, 1ris, 1sacA, 1sfp, 1smpI, 1stmA, 1svpA, 1tig, 1tiiD, 1tsg, 1tul, 1tupA, 1ubpA, 1uby, 1ulo, 1uroA, 1vcbA, 1wab, 1who, 1xbrA, 1ytbA, 1ytfD, 1zbdB, 2a0b, 2acy, 2bbkH, 2bgu, 2bpa2, 2cau, 2chsA, 2ezk, 2gmfA, 2hfh, 2ilk, 2mhr, 2occE, 2plc, 2pth, 2qwc, 2tbd, 3chbD, 3crd, 3lzt, 3pviA, 3ssi, and 3ullA.

The template selection in the threading portion of COBLATH depends on three parameters: the per-residue score f and the Z value of the top match and M, the number of structural neighbors of the top match that rank in the second to fifth places. Both true and false positives are expected to decrease as the selection criterion gets stricter. Figure 1 shows the dependence of the true and false positives on the per-residue score threshold f^* in qthread. At $f^* = 0.4$, the number of true positives declines sharply but the number of false positives stays at two. We therefore chose f^* to be 0.4. The Z^*_M values were chosen in a similar fashion.

The introduction of f and M in the template selection criterion is key to the large number of true positive assignments. In Figure 2 we plot the number of true positives as a function of error rate (false positive as percentage of total assignment) for the current selection criterion with varying f^* and a selection criterion based solely on the Z value. At comparable error rates, the latter criterion identified far fewer true positives. In particular, with $f^* = 0.4$ the current selection criterion identified 328 true positives at an error rate of 0.6% whereas a selection criterion with a Z value threshold of 6.5 would identify only 195 true positives at an error rate of 0.5%.

TABLE I. 390 Queries With Identified Templates and Alignment RMSDs and Sequence Identities

Query	Template	RMSD (Å)	Identity (%)	Query	Template	RMSD (Å)	Identity (%)
12asA	1b8aA	18.58	21	1iknD	1awcB	11.35	22
153l	1qsaA	8.28	22	1imbA	1inp	9.99	23
1a0fA	1gnwA	4.50	22	1ixh	1sbp	13.59	14
1a0i	1ckmA	8.76	17	1jfrA	1maaC	15.85	9
1a0p	1aihD	8.00	29	1jkmA	1maaC	20.08	15
1a28A	1errB	3.53	24	1jkw	1vin	13.78	17
1a3c	1bzyA	3.74	21	1jrhI	2hft	3.18	22
1a4iA	1psdA	14.36	11	1jxpA	1svpA	6.77	25
1a53	1pii	2.59	32	1kas	1pxtB	17.07	15
1a6m	1babA	1.75	25	1kb5B	1bec	3.69	30
1a7s	5ptp	2.20	30	1kpf	1guqA	4.52	19
1a7tA	1qh5B	7.40	19	1krs	1aszB	8.33	20
1a8h	1qu2A	4.74	19	1kte	1aazA	3.99	28
1aa0	1avyA	0.86	93	1kuh	1hfc	3.01	22
1aac	1ag6	4.50	26	1kum	1cyg	3.01	38
1afrA	1xikA	11.92	14	1lcl	1a3k	2.32	24
1afwA	1kas	8.09	18	1lfdA	1bt0	3.05	10
1agjA	1mctA	6.46	18	1lkkA	1d4tA	2.01	24
1ah1	1qfoA	6.53	15	1lrv	1bk5A	14.18	13
1ah7	1ca1	6.93	29	1lucA	1fvpA	4.63	18
1aj2	1rpxA	7.54	18	1lxa	2xat	15.57	17
1aj6	1yer	8.50	17	1mfmA	1yaiC	3.00	29
1ajsA	1bjwB	4.70	16	1mh1	5p21	1.97	34
1ak0	1ah7	5.97	17	1mjhA	5nul	11.58	10
1ak1	1qgoA	8.68	18	1moq	1ecfB	18.04	11
1ak4C	1eia	2.49	26	1msc	1qbeA	3.62	21
1ako	1bix	3.34	28	1mtyB	1mhyD	4.88	12
1alu	1rhgB	1.59	19	1mucA	1pdz	9.45	16
1am7A	1cnsA	8.43	14	1nar	2hvm	7.39	12
1amp	1xjo	5.94	24	1nbaA	1yacB	2.11	15
1amuA	1lci	7.81	19	1nbcA	1tf4A	8.25	26
1an8	3seb	3.89	23	1ndh	1qfzA	5.21	20
1an9A	1ojt	16.31	15	1neu	1cf8L	4.42	21
1aoiC	1hta	1.78	30	1nfn	1av1A	31.36	19
1apyB	1nedA	7.43	12	1nfp	1lucB	11.02	33
1aq0A	1edg	10.54	10	1nhp	1ebdA	7.49	24
1aqb	2a2uB	3.77	17	1nif	1aozB	7.54	18
1aquA	1nstA	9.13	16	1nkr	1wejL	12.51	10
1arv	1schA	8.82	18	1nksA	5tmp	5.63	24
1ash	2fal	2.05	12	1np4A	1avgI	7.47	18
1ass	1derA	3.62	20	1nsgB	1nfn	7.58	8
1at0	1am2	4.40	14	1nulA	1bzyA	5.56	18
1atg	1amf	2.04	25	1nwpA	1plc	3.20	26
1atiA	1qf6A	17.72	16	1oaa	1cydB	3.58	27
1atlA	1bkcA	6.83	28	1obpA	2a2uB	13.53	26
1atzA	1oakA	2.55	18	1ofgA	1gadP	10.42	15
1auoA	1maaC	14.12	12	1opr	1bzyA	10.84	11
1auq	1ido	3.74	17	1oyc	2tmdA	6.17	24
1auvA	2dln	8.37	24	1pbe	1lpfB	20.20	12
1avgI	1rbp	7.07	14	1pbwA	1rgp	3.20	17
1awd	2pia	2.81	30	1pdr	1qavA	1.45	37
1ax4A	1bjwB	8.05	13	1pea	2lbp	5.94	16
1axiB	1bquB	4.46	22	1pgtA	1axdB	3.35	26
1ayfA	1roe	9.44	17	1phd	1oxa	3.51	20
1aym1	1eah1	1.45	44	1phr	1tmy	9.67	13
1aym3	1qqp3	3.52	23	1plc	1bxvA	1.52	48
1ayx	1cem	12.66	14	1plq	1b77A	5.63	15
1b0nA	1rpeR	1.52	30	1poa	1faxL	12.17	16
1b20A	1rgeA	3.76	39	1pot	1anf	5.26	16
1b24A	1vdeB	6.53	20	1prxA	1qq2A	3.19	31
1b3rA	1psdB	14.03	19	1ps1A	5eau	7.16	16
1b4kA	1a53	17.22	11	1psrA	1bt6A	2.33	23

TABLE I. (Continued)

Query	Template	RMSD (Å)	Identity (%)	Query	Template	RMSD (Å)	Identity (%)
1b4vA	1gpeB	15.81	13	1pty	2shpA	2.07	35
1b51	1rh2F	2.80	57	1pysA	1lylA	19.23	15
1b8xA	1glqA	2.52	28	1qa9A	1cdcA	16.96	48
1b93A	1tmy	13.72	13	1qauA	1be9A	2.51	29
1b9dA	1cxqA	3.39	28	1qcxA	1air	12.70	23
1b9yC	1auc	2.23	21	1qddA	1bj3A	9.33	37
1ba1	1yagA	6.37	21	1qfoA	1a6wH	4.02	23
1b4vA	1gpeB	15.81	13	1pty	2shpA	2.07	35
1b51	1rh2F	2.80	57	1pysA	1lylA	19.23	15
1b8xA	1glqA	2.52	28	1qa9A	1cdcA	16.96	48
1b93A	1tmy	13.72	13	1qauA	1be9A	2.51	29
1b9dA	1cxqA	3.39	28	1qcxA	1air	12.70	23
1b9yC	1auc	2.23	21	1qddA	1bj3A	9.33	37
1ba1	1yagA	6.37	21	1qfoA	1a6wH	4.02	23
1bel	1reqA	4.92	28	1qqp2	2plv2	3.48	27
1bea	1hssD	5.18	34	1qreA	1lxa	14.84	16
1bebA	2a2uB	2.93	23	1rcf	5nll	2.40	25
1bfg	1hce	1.99	14	1rcy	1a65A	8.41	21
1bg6	91dtB	18.39	12	1rie	1ndoC	11.33	18
1bgc	2il6	3.91	16	1rlw	1rsy	2.39	23
1bgf	1bg1A	16.00	13	1rmg	1bhe	4.56	18
1bgp	1apxA	5.16	30	1rpxA	1gox	11.35	16
1bh5A	1han	15.99	19	1rsy	1djxB	3.90	29
1bjx	1a8y	6.73	20	1rtm1	1tn3	3.51	31
1bk0	1rxg	7.53	21	1ryc	1schA	8.72	17
1bk5A	3bct	6.67	16	1ryt	2fha	4.62	13
1bkrA	1aoa	2.71	19	1rzl	1hyp	5.38	28
1bli	1cyg	8.85	21	1sbp	1amf	6.17	17
1bmdA	5ldh	4.11	17	1shkA	3adk	4.71	17
1bncA	1iow	7.81	21	1smd	1cxlA	6.34	25
1bo4A	1nmtA	6.41	12	1smtA	6paxA	17.23	18
1boy	1fnhA	9.91	10	1smvA	2tbvB	18.59	22
1bquA	1axiB	4.31	22	1sra	2sas	10.75	9
1br0	1envA	26.68	15	1stfI	1cewI	5.09	24
1brt	1cqwA	8.05	19	1svy	1d0nB	2.05	36
1bt4A	1bj4A	7.33	15	1t1dA	1bax	9.56	18
1btkA	1qqgA	4.65	18	1taxA	7a3hA	6.50	12
1btl	1skf	5.26	19	1tbgA	1qksA	15.32	10
1btn	1pls	6.19	21	1tca	4lipE	18.42	13
1bv6	1bywA	4.40	11	1tcrA	2fgwL	5.05	26
1bxwA	1qj8A	3.20	21	1ten	1fnf	2.90	25
1by5A	1fepA	12.83	19	1tfe	1efuB	1.43	42
1byb	1b9zA	2.91	32	1tfr	1bgxT	13.58	20
1byi	2nipB	7.73	11	1theB	8pchA	3.91	33
1bykA	1dbqB	2.86	17	1thtA	1brt	12.50	14
1byw	1bv6	3.71	12	1tml	1qjwB	5.53	27
1c25	1rhs	11.43	18	1tx4A	1pbwA	3.19	17
1c3d	1ft1B	9.39	15	1tyfA	2dubD	9.77	17
1c9kB	1cydA	12.21	15	1uae	1eps	9.01	23
1cczA	1cdcA	21.01	21	1uok	1bvzA	4.97	28
1cd1A	1hsaD	4.28	22	1urnA	2sxl	4.60	24
1cd8	1bj1L	3.02	26	1vcaA	1iam	6.18	26
1cdkA	1a06	11.78	35	1vdrA	1drf	3.42	23
1cdy	1cf8L	12.87	21	1vfrA	1nox	2.15	26
1ceo	1edg	3.59	19	1vhrA	1rpmB	13.32	12
1ceqA	2cmd	3.96	27	1vid	2admB	4.40	14
1cewI	1stfI	2.63	21	1vin	1bu2A	2.60	21
1cex	1bs9	3.49	22	1vls	1cpq	4.53	15
1cfb	1mfn	11.31	18	1vmoA	9wgaA	12.07	11
1cg2A	1amp	10.56	17	1vpfA	1pdgC	2.41	25
1chmA	1az9	5.08	21	1wba	1avaC	4.74	20
1cid	1cdy	11.35	28	1whtB	1ivyB	1.62	23

TABLE I. (Continued)

Query	Template	RMSD (Å)	Identity (%)	Query	Template	RMSD (Å)	Identity (%)
1ckmA	1a0i	10.58	18	1wit	1tlk	2.32	21
1c12A	1bjwB	14.28	15	1xel	1bxkA	6.69	22
1cnt3	1bgc	2.43	23	1xikA	1xsm	2.52	24
1cnv	1ctn	10.46	13	1xjo	1amp	3.06	23
1cof	1ak7	2.94	35	1yacA	1nbaB	5.74	16
1cp2A	1fts	10.43	14	1ybvA	1ae1B	2.72	26
1cpcA	1cpcL	2.28	22	1ycc	1c6s	4.34	21
1csn	1a06	7.44	18	1ycsB	1awcB	2.11	26
1cto	1bj8	7.30	17	1yer	1b63A	6.97	26
1cviA	1ihp	7.43	20	1zin	1bif	10.57	14
1cvl	1brt	10.89	28	1zpdA	1poxB	7.08	15
1cyx	2occB	4.25	29	1zxq	1vcaA	5.36	21
1d2nA	1a5t	12.39	13	256bA	1bbhA	3.92	28
1dapA	1ofgA	17.65	11	2abk	1mun	3.58	20
1ddf	1fadA	2.81	24	2cbp	1nwpA	5.46	33
1dhn	1b66A	6.20	17	2ccyA	1cgn	3.37	32
1dhpA	1nal1	2.04	24	2dorA	1gox	14.22	20
1dhr	1ybvA	7.01	19	2dri	1dbqB	3.44	23
1dosA	2tysA	6.17	13	2ebn	1ctn	17.45	12
1dpsA	1bcfB	4.23	19	2ercA	2admB	3.39	21
1drw	1gadP	16.53	13	2fha	1bcfB	2.20	21
1dssG	1dapA	19.95	11	2gar	1fmtB	9.33	16
1dun	1dupA	3.05	20	2gdm	1babB	3.21	17
1dupA	1dutB	1.62	32	2gsaA	2oatA	3.79	27
1dxy	1psdB	2.85	26	2hbg	1mba	5.05	22
1dynA	1pls	2.90	21	2ilb	2ila	3.86	25
1eaf	3cla	4.18	27	2lbd	1lbd	7.91	30
1eceA	7a3hA	4.56	17	2liv	1pea	6.11	15
1ecpA	1cb0A	6.06	14	2mbr	1qltB	12.53	18
1edg	1tr1C	10.52	13	2mcm	1xbd	3.93	34
1edhA	1fnhA	14.51	13	2mev1	1qqp1	9.98	22
1eerB	1hwgC	5.50	18	2mprA	1a0tP	7.90	28
1efvA	1efpD	11.12	17	2mtaC	1c6s	3.18	16
1eny	1enp	3.52	31	2nadA	1psdB	4.98	25
1erv	1xob	1.69	28	2nmbA	1shcA	3.75	25
1etb1	1bpv	6.83	0	2omf	3prn	7.43	15
1exg	2xbd	2.22	27	2pia	1cqxB	5.78	27
1exnA	1bgxT	10.92	21	2polA	1plq	3.79	15
1fcdA	1ebdA	8.89	14	2por	1pho	13.13	16
1fdr	2cnd	3.55	15	2pvbA	1ahr	13.63	26
1fepA	2fcpA	10.94	18	2rspB	1vikA	9.29	25
1fgs	1uag	5.42	16	2sak	1bt0	10.33	23
1fit	4rhn	3.10	22	2scpA	4cln	15.48	19
1flp	1mba	2.28	25	2tbvA	4sbvB	5.80	27
1fltX	3ncmA	2.26	28	2tct	2ktqA	21.90	16
1fmb	1hvc	2.72	30	2tgi	1agqC	2.89	18
1fnc	1amoB	2.41	27	2tpsA	1rpxC	4.81	16
1fnf	1fnhA	9.96	31	2tysA	1rpxC	4.61	14
1frb	1qrqA	3.00	22	2vhbA	1cqxB	2.76	51
1frpA	1imbA	8.69	16	2vil	1d0nB	2.89	61
1ft1A	1a17	9.72	20	3chy	1tmy	1.86	29
1ft1B	2sqcA	16.98	18	3cla	1eaf	12.54	25
1furA	1dcnC	4.14	19	3grs	1lvl	2.86	26
1fyc	1lac	3.49	31	3inkC	1alu	6.25	21
1gc1H	1igtB	2.02	53	3nul	1pne	3.06	27
1gen	1fbl	2.53	36	3pfk	2dri	11.38	19
1gky	1zin	8.11	17	3pte	2blsB	4.04	24
1gox	1b3oB	2.04	22	3pyp	1byw	9.67	10
1gr2A	1wdnA	5.53	30	3sdhA	1babA	2.19	18
1gsa	1bncB	7.74	11	3seb	1an8	5.96	25
1gsoA	1bncB	6.83	18	3sil	1eur	5.35	22
1guxB	1tfb	7.35	20	3thi	4mbp	5.00	18

TABLE I. 390 Queries With Identified Templates and Alignment RMSDs and Sequence Identities

Query	Template	RMSD (Å)	Identity (%)	Query	Template	RMSD (Å)	Identity (%)
1han	1mpyB	5.21	20	3tmkA	5mpA	3.43	25
1hfc	1sat	6.52	33	4crxA	1a0p	13.70	13
1hlb	1gbuD	2.89	23	4icb	1sra	3.19	19
1hqi	2mobA	6.81	24	4mbp	3thi	4.93	19
1huw	1au1B	14.82	12	4xis	1a0cC	3.72	26
1hxn	1pex	4.83	24	5nul	1moB	3.67	15
1iakA	1biiA	9.14	25	5ptp	1a7s	2.06	31
1iakB	1zagC	3.96	29	7a3hA	1bqcA	4.49	18
1ido	1aoxB	3.59	26	8abp	1gca	3.54	20
1igtB	35c8H	2.61	69	8fabA	1wejH	5.54	31
1ihp	1rpa	7.74	19	9rnt	1aqzA	4.02	28

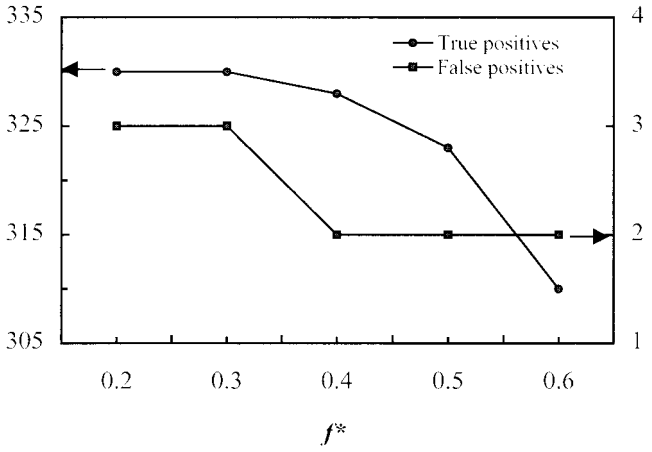

Fig. 1. The dependence of the number of true and false positive templates (assigned by qthread for 533 FSSP proteins) on the threshold of the per-residue score of the top match. The threshold of the *Z* value was 5.0, 4.0, 4.0, 4.0, or 2.0 when the number of structural neighbors of the top match ranking at the second to fifth places was 0 to 4.

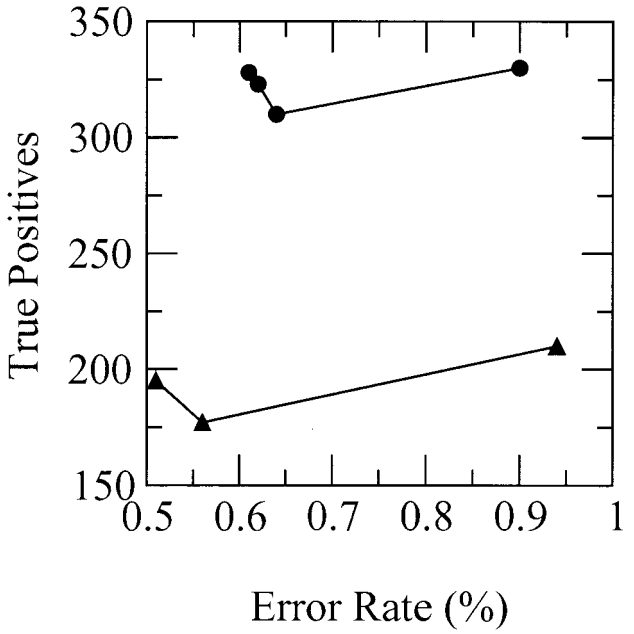

Fig. 2. The number of true positives in qthread of 533 FSSP proteins plotted against the error rate (false positive as percentage of total assignment). Two selection criteria are compared. For the upper curve the selection uses both a threshold of the per-residue score and an *M*-dependent threshold of the *Z* value. From left to right the four points correspond to $f^* = 0.4$, 0.5, 0.6, and 0.3, respectively (Z_M^* kept at the values in Fig. 1). For the lower curve the selection uses a universal threshold of the *Z* value. From left to right the three points correspond to the threshold set at 6.5, 7.0, and 6.0, respectively.

How important are the predicted secondary structure and solvent exposure in the threading? If only the sequence profiles were used in the scoring function, the number of queries assigned templates by qthread was 279 (with one false positive). Thus the inclusion of the predicted secondary structure and solvent exposure allowed for an additional 51 template assignments (to a total of 330).

We also tested the importance of the sequence profile by replacing it with a position-independent residue substitution matrix. For this we chose the Gonnet matrix,[41] which had been shown to perform somewhat better than the BLOSUM62 matrix in threading.[36] Only 270 queries were assigned templates (with three false positives). This outcome is even worse than that obtained by using the sequence profile alone, clearly illustrating the importance of the position-specific sequence profile from PSI-BLAST in discriminating the true template from decoys.

Alignment Accuracy

The alignment accuracy of COBLATH was assessed by the RMSD between the predicted and actual C_α positions. In particular, we were interested in the alignments that had RMSDs < 8 Å. Beyond that cutoff, the alignment is probably not useful for building a reasonable structural model for the query.

For the 307 templates identified by qblast, 202 query-template alignments had RMSDs < 8 Å. Remarkably, qthread by the local-local algorithm was found to improve the alignment accuracy. For the same 307 query-template pairs, the number of alignments with RMSDs < 8 Å increased to 223. The RMSDs of 50 qthread alignments were lower by 20%; only about half that many, i.e., 28, saw their RMSDs increase by 20%. We thus took the protocol of threading by the local-local algorithm as our choice for

query-template alignment, regardless of whether the template was identified by PSI-BLAST or threading. The threading is either qthread or tthread, depending on whether the template is identified by qblast/qthread or tblast/tthread. By this protocol, 265 query-template alignments (68% of all 390 identified) had RMSDs < 8 Å. Of these, a majority (191, i.e., 71%) had RMSDs within 4 Å.

The accuracy of the sequence-structure alignments can also be assessed by comparing against structure-structure alignments. Of the 390 query-template alignments obtained by the threading, 281 (or 72%) have more than 50% of the aligned residues identical to those obtained by structure-structure alignment using the DALI score and 297 (or 76%) have more than 65% of residues aligned to within four positions of the structure-structure alignments. Of the 265 threading alignments with RMSDs < 8 Å , 28 had worse than 50% agreement with structural alignments. In comparison, 44 threading alignments had more than 50% agreement with structural alignments and yet had RMSDs > 8 Å. The RMSD < 8 Å criterion thus appears to be somewhat stricter than the criterion of $>$ 50% agreement with structural alignment.

Among the 390 query-template alignments, 174, 165, and 36 had sequence identities of 10–19%, 20–29%, 30–39%, respectively. Of the remaining 15 alignments, five had sequence identities $< 10\%$ whereas 10 had sequence identities $\geq 40\%$. The RMSDs and sequence identities of the 390 query-template alignments are listed in Table I.

The above level of alignment accuracy was achieved by using reduced gap penalty in the threading by the local-local algorithm (5 for gap opening and 0.3 for gap extension). If the gap penalty remained at what was used for the fold recognition (twice the above values), only 243 (compared to 265) of the query-template alignments have RMSDs < 8 Å. The number of alignments with more than 50% aligned residues identical to those of structure-structure alignments reduced from 281 to 267. For the 307 templates identified by qblast, using the stronger gap penalty reduced the number of alignments with RMSDs $<$ 8 Å from 223 to 206, which is now almost the same as the number (202) obtained by PSI-BLAST.

Results on a Set of 68 Proteins

The set of 68 proteins compiled by Fischer and Eisenberg[36] is a standard for testing fold recognition methods. The goal of the original work was to identify templates from a library of selected 301 folds; specifically, to see whether the intended templates were ranked at the top. As such, only the qblast and qthread methods of COBLATH were appropriate for this set of queries. For 32 queries, the intended templates were found by qblast and had the lowest h values (sometimes after excluding homologues of the intended templates). Qthread performed exceptionally well, ranking 56 intended templates at the top. The success rate is thus 82%. In comparison, the best method evaluated by Fischer and Eisenberg had a success rate of 76% (http://www.doe-mbi.ucla.edu/people/fischer/BENCH/table1.html).

Jones[19] also studied the set of 68 proteins, examining not only the ranks of the intended targets but also the query-template alignment accuracy. He found that, for 22 queries, the alignments by his threading method were in agreement with structure-structure alignments over more than 50% of the aligned residues. In comparison, the protocol of qthread by the local–local algorithm yielded twice as many, i.e., 43, sequence–structure alignments that had more than 50% agreement with structure–structure alignments. Thirty-six qthread alignments had RMSDs < 8 Å; of these only four had worse than 50% agreement with structural alignments. In comparison, 13 qthread alignments had more than 50% agreement with structural alignments and yet had RMSDs > 8 Å. Again the RMSD < 8 Å criterion appears stricter than the criterion of $> 50\%$ agreement with structural alignment. The results of our study on the 68 queries are summarized in Table II, which lists the ranking of the intended template, the RMSD of the query–template alignment, and the agreement between threading and structural alignments (as percentage of identically aligned residues).

Accuracy of 12 Query–Template Alignments

Recently Kolinski et al.[40] (KRIS) developed a sophisticated method that they found to improve the alignment accuracy of threading. Actually 10 of the 12 query–template pairs were identical to those in the set compiled by Fischer and Eisenberg. The two additional query–template pairs were 256bA with 1bbhA and 2pcy with 2azaA. We thought it would be interesting to test the protocol of qthread by the local–local algorithm against a method specifically designed for alignment accuracy.

The RMSDs of the alignments by the method of KRIS and by our qthread protocol are compared in Table III. Qthread gave lower RMSDs for seven queries and higher RMSDs for the other five. The algebraic mean of the 12 RMSDs is 7.2 Å by the method of KRIS and 6.4 Å by our qthread protocol. The qthread protocol performs as well as the method of KRIS.

Annotation of the MG Genome

The number of MG ORFs for which templates were assigned was 224, 208, 218, and 175 by qblast, tblast, qthread, and tthread, respectively. Significantly, each method identified templates for a large number of ORFs that did not have templates assigned by other methods. Specifically, 47 of the ORFs were assigned templates by qblast but not by tblast, and 31 ORFs were assigned templates by qblast but not by qblast. By merging the results of qblast and tblast, the total number of ORFs assigned templates by two-way PSI-BLAST was 255. Similarly, 56 of the ORFs were assigned templates by qthread but not by tthread, and 13 ORFs were assigned templates by qthread but not by qthread. By merging the results of qthread and tthread, the total number of ORFs assigned templates by two-way threading was 231. Of the 225 ORFs not assigned templates by two-way PSI-BLAST, 43 were assigned templates by two-way threading.

The consistency of the four methods was checked in cases where more than one method identified templates for

TABLE II. Ranking of 68 True Templates and Query-Template Alignment Accuracy

Query	Template	Rank	RMSD (Å)	Agreement (%)
1aaj	1paz	1	5.36	68
1aba	1ego	1	3.38	66
1aep	256bA	1	13.66	0
1arb	5ptp	1	12.36	23
1atnA	1atr	1	8.04	49
1bbhA	2ccyA	1	3.51	83
1bbt1	2plv1	1	12.27	59
1bgeB	2gmfA	12	8.01	22
1c2rA	1ycc	1	3.00	89
1cauB	1cauA	1	3.65	77
1cewI	1molA	1	13.20	43
1chrA	2mnr	1	3.02	79
1cid	2rhe	18	12.31	48
1cpcL	1colA	1	17.06	0
1crl	1ede	1	17.65	15
1dsbA	2trxA	3	5.28	52
1dxtB	1hbg	1	1.96	96
1eaf	4cla	1	4.18	83
1fc1A	2fb4H	1	8.26	63
1fxiA	1ubq	1	10.76	30
1gal	3cox	1	13.75	51
1gky	3adk	1	7.88	38
1gp1A	2trxA	5	9.30	57
1hip	2hipA	1	3.86	81
1hom	1lfb	1	4.80	90
1hrhA	1rnh	1	3.91	75
1isuA	2hipA	2	2.78	75
1lgaA	2cyp	1	3.54	74
1ltsD	1bovA	11	9.58	0
1mdc	1ifc	1	1.92	98
1mioC	2minB	1	13.07	69
1mup	1rbp	1	7.24	64
1npx	3grs	1	7.08	66
1onc	7rsa	1	4.34	77
1osa	4cpv	1	15.55	72
1pfc	3hlaB	1	3.69	84
1rcb	2gmfA	1	7.14	33
1sacA	2ayh	1	14.51	0
1stfI	1molA	1	12.98	37
1tahA	1tca	1	10.08	82
1ten	3hhrB	1	2.72	79
1tie	4fgf	10	9.90	0
1tlk	2rhe	1	5.04	59
2afnA	1aozA	1	8.68	46
2ak3A	1gky	1	15.59	30
2azaA	1paz	8	3.82	42
2cmd	6ldh	1	4.21	76
2fbjL	8fabB	1	3.17	84
2gbp	2liv	1	12.56	54
2hhmA	1fbpA	1	7.45	59
2hpdA	2cpp	1	5.04	74
2mnr	4enl	1	8.36	57
2mtaC	1ycc	1	5.25	50
2omf	2por	1	10.43	30
2pia	1fnb	1	10.61	66
2pna	1shaA	1	3.67	85
2sarA	9rnt	1	3.80	55
2sas	2scpA	1	3.59	96
2sga	5ptp	1	10.63	50
2sim	1nsbA	5	9.93	42
2snv	5ptp	4	11.51	38
3cd4	2rhe	1	8.50	52
3chy	2fox	2	4.41	0
3hlaB	2rhe	1	8.84	73
3rubL	6xia	23	10.69	0
4sbvA	2tbvA	1	6.54	78
5fd1	2fxb	1	9.61	13
8i1b	4fgf	1	10.65	33

TABLE III. C_α RMSDs (in Å) of 12 Query-Template Alignments by Two Methods

Query	Template	KRIS	Qthread
1aba	1ego	4.86	3.38
1bbhA	2ccyA	6.82	3.51
1cewI	1molA	14.38	13.20
1hom	1lfb	3.70	4.80
1stfI	1molA	5.95	12.98
1tlk	2rhe	4.17	5.04
256bA	1bbh	4.36	3.92
2azaA	1paz	10.77	3.82
2pcy	2azaA	4.41	5.65
2sarA	9rnt	7.83	3.80
3cd4	2rhe	6.39	8.50
5fd1	2fxd	12.40	9.61
Mean RMSD		7.2	6.4

the same ORF. For each of the 177 ORFs which had templates assigned by both qblast and tblast, the templates were either identical or were structural neighbors according to the FSSP library. Similarly, For each of the 162 ORFs which had templates assigned by both qthread and tthread, the templates were either identical or were structural neighbors according to the FSSP library. However, for the 188 ORFs which had templates assigned by both two-way PSI-BLAST and two-way threading, conflict arose in a single case. The ORF MG393 was aligned with 1aonO by two-way PSI-BLAST but was aligned with 1iyu by qthread. The Z value of threading MG393 to 1iyu was 4.1, just barely passing the threshold of $Z_2^* = 4.0$. 1aonO was taken to be the correct template.

Overall COBLATH identified templates for 298 of the 479 MG ORFs. Of these, 35 were likely to be transmembrane proteins and 11 were likely to have coiled-coil structures. The fraction of MG ORFs aligned with globular templates is thus 53%. This is higher by 10–20 percentage points than those from the current literature (see Table IV). The 298 MG ORFs for which templates have been assigned can be viewed from our web page at http://cmbph1.physics.drexel.edu/MG/MG.html.

Teichmann et al.[29] identified templates for 213 ORFs by two-way PSI-BLAST. Of these 207 ORFs were identified by COBLATH. Only two of the 207 ORFs with templates assigned both by COBLATH and by Teichmann et al., MG356 and MG397, had conflicting assignments. Teichmann et al. assigned both ORFs to 1bgw whereas we assigned the former to 1fgkA and the latter to 1bk5A. Our template assignment for MG356 was consistent with that by Huynen et al.[28] The six ORFs assigned templates by Teichmann et al. but not by COBLATH were: MG130, MG140, MG141, MG312, MG353, and MG468. In comparison, 47 ORFs were assigned globular templates by COBLATH but not by Teichmann et al. Twenty-three of these have been assigned templates by other groups listed in Table IV. The remaining 24 new template assignments are listed in Table V. Of these, seven (for MG027, MG207, MG232, MG246, MG293, MG311, and MG434) could be attributed to the fact that the template structures were

TABLE IV. Annotation Results of the MG and SC Genomes From Previous and the Present Studies

	MG		
Authors (year)	Method	Annotated fraction (%)	Web page
Fischer & Eisenberg (1998)	Threading	35	www.doe-mbi.ucla.edu/people/Frsvr/preds/MG/MG.html
Huynen et al. (1989)	PSI-BLAST	37	dove.EMBL-Heidelberg.de/3D/MG.pred
Rychlewski et al. (1998)	Threading	34	bioinformatics.ljcrf.edu/FFAS_genomes/genomes.html
Teichmann et al. (1999)	Two-way PSi-BLAST	44	www.mrc-lmb.cam.ac.uk/genomes/MG
Jones (1999)	Threading	40	globin.bio.warwick.ac.uk/genome/genomedb.cgi
Wolf et al. (1999)	PSI-BLAST	34	ncbi.nlm.nih.gov/pub/koonin/FOLDS/genometable.html
Gerstein (1999)	Fasta + PSI-BLAST	34	bioinfo.mbb.yale.edu/genome/db99Mgen/structure_matches.txt
Muller et al. (1999)	PSI-BLAST	28	www.bmm.icnet.uk/PsiBench/html/tbmg_main.html
Frishman (2000)	PSI-BLAST	33	pedant.mips.biochem.mpg.de
COBLATH (1999)	PSI-BLAST + Threading	62	cmbph1.physics.drexel.edu/MG/MG.html
	SC		
Authors (year)	Method	Annotated fraction (%)	Web page
Sanchez & Sali (1998)	Threading	36	pipe.rockfeller.edu/modbase
Wolf et al. (1999)	PSI-BLAST	22	ncbi.nlm.nih.gov/pub/koonin/FOLDS/genometable.html
Hegyi & Gerstein (1999)	Fasta + PSI-BLAST	27	bioinfo.mbb.yale.edu/genome/db99 Scer/structure_matches.txt
Elofsson & Sonnhammer (1999)	Hidden Markov Models	22	www.biokemi.su.se/research/Elofsson-Arne.html
Frishman (2000	PSI-BLAST	25	pedant.mips.biochem.mpg.de
Jones (1999)	Threading	34	globin.bio.warwick.ac.uk/genome/genomedb.cgi
COBLATH (1999)	PSI-BLAST + Threading	45	cmbph1.physics.drexel.edu/yeast/yeast.html

released after the other studies. Most (19) of the 24 template assignments were made only by the threading.

For 15 of the 298 MG ORFs, two templates were assigned to different regions of their sequences. The 298 ORFs were thus mapped to a total of 313 templates. These cover 71.2% of the 109,862 residues of the 298 ORFs. The remaining 181 ORFs have 64,704 residues. Thus in total the structural annotation by COBLATH covered 44.8% of the residues of the MG genome.

Of the 313 ORF-template alignments, those with sequence identities in 0–9%, 10–19%, 20–29%, 30–39%, and 40–79% numbered 13, 127, 72, 64, and 37, respectively. The 10–29% block represented 64% of all the template assignments.

The 313 templates involve only 195 unique FSSP proteins. If mapping to the same FSSP proteins implies originating from the same gene, then gene duplication had occurred in 313–195 = 118 cases, or 37.7% of the 313 MG ORF regions. The 195 FSSP proteins represent 10% of the full library of 1907 FSSP proteins used in the present study.

Of the 313 templates, those in the α, β, and mixed classes were 24, 4, and 72%, respectively. The class assignments of the ORFs according to our secondary structure prediction agreed with those of the templates in 86% of the 313 cases.

Annotation of the SC Genome

By qblast, 2113 ORFs were assigned templates. Qblast produced template assignments for 386 new ORFs, and two-way threading gave assignments to additional 384 ORFs. Of the last 384 ORFs, 136 were given the same template assignments by both qthread and tthread. Only one ORF, YDR289C, was given conflicting assignments by the two threading methods (the assignments were rejected). The consistency between qthread and tthread indicates their reliability.

In total, 2113 + 386 + 384 = 2883 of the 6337 SC ORFs were assigned templates. Of these, 239 were likely to be transmembrane proteins and 46 were likely to have coiled-coil structures. The fraction of SC ORFs aligned with globular templates is thus 41%. This is higher by 5–20 percentage points than those from the recent literature (see Table IV). The 2883 SC ORFs with template assignments can be viewed from our web page at http://cmbph1.physics.drexel.edu/yeast/yeast.html.

TABLE V. Novel Template Assignments of 24 MG ORFs

ORF	ORF description	Template	PDB description
MG010	—	2dri	D-ribose-binding protein
MG018	ATP-dependent RNA helicase	1heiA	hcv helicase (HELICASE)
MG022	DNA-directed RNA polymerase	1a8y	calsequestrin
MG027	—	1qbzA	siv gp41 ectodomain fragment mutant
MG060	—	1xel	udp-galactose 4-epimerase
MG073	excinuclease ABC subunit B (uvrB)	1heiA	hcv helicase (HELICASE)
MG150	ribosomal protein S10	1ris	ribosomal protein s6
MG158	ribosomal protein L16	2nmbA	numb protein fragment gppy peptide
MG207	—	1uteA	ii purple acid phosphatase
MG232	ribosomal protein L21	2cuaA	cua fragment
MG246	—	1ush_	5′-nucleotidase (udp-sugar hydrolase)
MG252	rRNA methylase	2liv	L/I/V-binding protein
MG258	peptide chain release factor 1	1sesA	seryl-trna synthetase
MG284	—	1psrA	psoriasin
MG293	phosphodiesterase	1qumA	endonuclease iv
MG299	phosphotransacetylase	1fsz	ftsz (sulb)
MG311	ribosomal protein S4	1c05A	ribosomal protein s4 delta 41 fragment
MG335.1	—	1akhA	a-1 mating-type protein alpha-2
MG368	fatty acid synthesis protein	3pfk	phosphofructokinase
MG396	ribose-5-phosphate isomerase	1ntr	ntrc receiver domain
MG426	ribosomal protein L28	1azpA	sac7d (7 kd DNA-binding protein)
MG434	uridylate kinase (pyrH)	2scuA	succinyl-coa ligase (scs)
MG445	tRNA guanine-N1-methyltransferase	2bgu	beta-glucosyltransferase
MG454	—	1rl6A	ribosomal protein 16 biological unit

The web page of Jones lists template assignments for 2107 ORFs. Of these, 1936 are among the 2883 ORFs for which COBLATH have assigned templates. The two sets of template assignments were consistent in all but 29 cases. Jones assigned templates for 171 ORFs not annotated by COBLATH, whereas COBLATH assigned templates for 947 ORFs not annotated by Jones. Excluding those (250) likely to be transmembrane proteins or likely to have coiled-coil structures and those (185) assigned to templates with structures released after the work of Jones, COBLATH assigned templates for 512−171 = 341 more ORFs than Jones. The difference is 5% of the whole genome.

For 262 of the 2883 SC ORFs, two templates were assigned to different regions of their sequences. Another 48 ORFs had three or more separate regions assigned templates. The 2883 ORFs were thus mapped to a total of 3266 templates. These cover 50.0% of the 1,648,420 residues of the 2883 ORFs. The remaining 3454 ORFs have 1,335,889 residues. Thus in total the structural annotation by COBLATH covered 27.6% of the residues of the SC genome.

Of the 3266 ORF-template alignments, those with sequence identities in 0–9%, 10–19%, 20–29%, 30–39%, 40–69%, and 70–100% numbered 114, 1329, 1033, 380, 323, and 87, respectively. The 10–29% block represented 72% of all the template assignments.

The 3266 templates involve only 725 unique FSSP proteins. If mapping to the same FSSP proteins implies originating from the same gene, then gene duplication had occurred in 3266−725 = 2541 cases, or 77.8% of the 3266 MG ORF regions. This rate of gene duplication is more than double that in the MG genome.

The 725 FSSP proteins represent 38% of the full library of 1907 FSSP proteins used in the present study. They include 150 of the 195 FSSP proteins used in structural annotation for the MG genome. In other words, the structural annotation of the MG genome only involved 45 FSSP proteins not used for the SC genome, but the structural annotation of the SC genome involved 575 FSSP proteins not used for the MG genome.

Of the 3266 templates, those in the α, β, and mixed classes were 30, 9, and 61%, respectively. The class assignments of the ORFs according to our secondary structure prediction agreed with those of the templates in 80% of the 3266 cases.

DISCUSSION

Advantage of COBLATH

We have demonstrated the complementarity of two existing fold recognition methods, threading and PSI-BLAST. The complementarity appears not just as enhanced success rate of fold recognition. The accuracy of the resulting alignment is found to be higher than found in previous studies.

Various ingredients of COBLATH were previously found to be very useful for fold recognition. Two-way PSI-BLAST have been shown to be more effective in fold recognition than the usual protocol of using the query sequence as search input.[31] We further improved two-way PSI-BLAST by including the structural alignments of each potential template as part of the input when searching for queries from the template sequence. Sequence profiles from PSI-BLAST have been found to improve the ability of threading in fold recognition.[28] Here they are combined with a substitution matrix involving the secondary structure

predicted by arguably the best available method. Furthermore, we used the predicted secondary structure to screen the fold library so a more focused search can be made by threading. Following Rychlewski et al.,[14] we also implemented the idea of two-way searching in threading. In essence, we have integrated some of the most useful ingredients into COBLATH.

Factors Affecting the Performance of Threading

The position-specific sequence profile from PSI-BLAST was found to be far better than a position-independent residue substitution matrix for fold recognition by threading. It is perhaps the most important among all contributing factors in discriminating the true template from decoys. Predicted secondary structure and solvent exposure carry significant information for fold recognition. In the present study, it was found to be responsible for identifying as much as 15% of templates (increasing the number of template assignments from 279 to 330 in qthread).

The criterion for deciding whether the top match in fold recognition by threading should be identified as the template affects the performance of threading in a fundamental way. Substantial improvement is achieved through the introduction of the number M of the top match's structural neighbors ranking the second to fifth places. Obviously, if nonhomologous structural neighbors are also ranked among the best, then the probability that the top match gets there by chance is reduced. As such, a lower threshold of the Z value for selecting the top match can be used. The lowered threshold is responsible for the superiority of the current selection criterion (as compared with a conventional criterion based solely on the Z value). Indeed about half of the templates identified by our threading had all the next four places occupied by structural neighbors and were thus eligible for using a low Z threshold. For example, 158 of the 330 templates identified for the 533 FSSP proteins by qthread had $M = 4$. A threshold of the per-residue score f of the top match helps screen out false positives.

A threading protocol suitable for fold recognition may not be ideal for achieving optimal alignment between query and template. Indeed we were able to improve the alignment accuracy by switching to a local–local algorithm and reduced gap penalty. Both were designed to focus the alignment to regions where query and template sequences are more similar. The improvement allowed threading to outperform PSI-BLAST in alignment accuracy. As a result, we have selected threading (by the local–local algorithm and with reduced gap penalty) as our final choice for generating query–template alignments. In short, the utility of the threading portion of COBLATH is twofold. It allows additional queries to be assigned templates and provides better alignments for all query–template pairs.

Further Improvement

The obvious question is what about the queries not assigned templates by COBLATH (e.g., the remaining 143 FSSP proteins in the test set of 533). Typically the structural neighbors of these queries were either very close homologues (with identities higher than 90%) or very remote homologues (sometimes known as analogs), with alignments characterized by many large gaps and low identities. Take, for example, the query 1ulo (with 152 residues). The first structural neighbor of 1ulo in FSSP is 1ulp, which has 100% sequence identity. The next structural neighbor is 1ciy, having 114 of its 577 residues aligned to 1ulo with an RMSD of 3.1 Å and sequence identity of just 9%. The remaining neighbors have equally poor or worse structural alignments. Additional sequences and structures deposited to databases will undoubtedly improve the performance of COBLATH and other methods. Additional sequences can help if they fill in the void between close and remote homologues. Additional structures, even those belonging to one of the folds in the current FSSP library, will increase the chance of hitting a template.

Other factors such as contact energy have been found to be useful for fold recognition.[28] It would be interesting to successively include such factors in the threading scoring function to search for additional discriminatory power. Combining sequence profiles of structural neighbors may also prove useful.[42]

In the current version of COBLATH, improvement may come from refining the template selection criteria of qthread and tthread. The criteria must be designed such that as few false positives as possible are included. In doing so, a number of true positives may be rejected. For example, by qthread of the 533 FSSP proteins, true templates were ranked at the top in 402 cases. However, the current criterion picks out only 328 true templates (plus two false positive). The selection of the remaining 74 true templates may be possible by a refined criterion or by producing higher matching scores.

One may wonder whether the screening, which reduces the number of potential targets from 1907 to just 100, adversely affects the success rate of fold recognition. The answer is basically no. First of all, the screening is quite accurate, rejecting the queries themselves only in 49 of 533, or 9% of cases. Secondly, even among these 49 queries, COBLATH correctly identified templates in 21 cases. Of course, a better secondary-structure prediction method, especially for β proteins, will improve both the screening and the matching scores in threading. The screening makes it practical to apply COBLATH to large-scale structure prediction, for example, on the SC genome and genomes of higher organisms.

How Long Away Are We From Complete Annotation of the Two Genomes

With 62% of the MG genome annotated, it is intriguing to speculate how long will it take for the remaining ORFs to be structurally annotated if structure determination is kept at the current pace.

The FSSP library in the present study is the November 21, 1999, edition and has 1907 entries. The edition of April 23, 2000, has grown to 2144 entries. Hence in the intervening five months structure determination has yielded 237 new entries. This corresponds to 569 new entries annually.

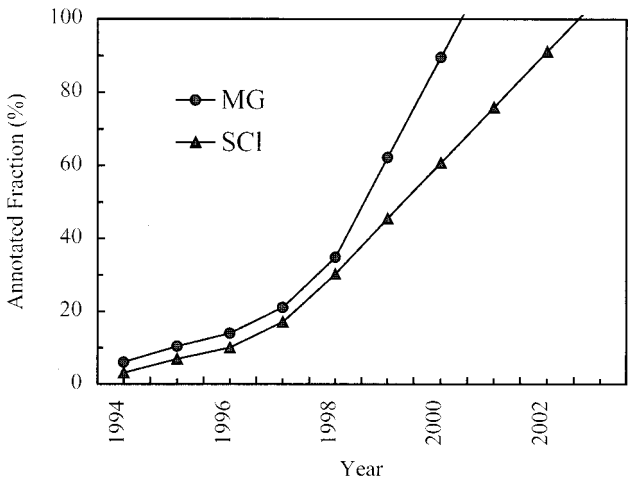

Fig. 3. The fraction of MG and SC ORFs that would have been structurally annotated in a particular year if only PDBs released in that year were available. Points after 1999 are extrapolated from the points at 1998 and 1999.

If, like for the 298 MG ORFs that have been annotated, 10% of all FSSP proteins will be used for the annotation of the remaining MG ORFs, then each year 57 of the annual production of FSSP proteins will become templates of the remaining ORFs. Since the annotation of the 298 ORFs involved 195 FSSP proteins, the remaining 181 ORFs are expected to require $195 \times 181/298 = 118$ new FSSP proteins. This corresponds to $118/57 = 2.1$ years of structure determination. In two years' time, with the current COBLATH method, we expect the MG genome will be completely annotated.

A similar estimate can be made for the complete annotation of the SC genome. The annotation of the 2883 ORFs took 725 FSSP proteins, hence the remaining 3454 ORFs are expected to require $725 \times 3454/2883 = 869$ new FSSP proteins. Annually we expect to have $569 \times 38\% = 216$ new FSSP proteins that will be templates of the remaining ORFs. Thus the complete annotation of the SC genome will take another $869/216 = 4$ years.

These estimates are equivalent to treating the ORFs assigned to all templates with structures released in a particular year as that year's annotation yield and then extrapolating to future years so the full genome is covered. Figure 3 displays such an extrapolation. Plots of this type have been created previously. The estimate of Gerstein and Hegyi[43] was fairly pessimistic—reaching 100% annotation of the MG genome in the year 2050. A recent analysis[44] led to a more optimistic view, but perhaps because the slope of the curve near the end was almost flat, the authors did not venture an estimate. In that analysis, annotation results from several studies were pooled in order to increase the annotated fraction. Structurally similar templates can have very different release dates, so an ORF can be assigned to different years. Since different studies have different ways of selecting a template for a given ORF, the pooling of results from different studies tends to randomize the date-allocation process. This may explain the relatively flat slope.

A Powerful Tool for Structural Genomics

COBLATH has a 73% success rate of fold recognition for the 533 FSSP proteins. Of the 390 query–template alignments predicted, 68% have RMSDs < 8 Å. As illustrated by the annotation of the MG and SC genomes, COBLATH may prove to be a powerful tool for structural genomics. The COBLATH server is at http://cmbph4.physics.drexel.edu/COBLATH/submit.html.

REFERENCES

1. Bowie JW, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
2. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
3. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci USA 1992;89:12098–12102.
4. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
5. Flockner H, Braxenthaler M, Lackner P, Jaritz M, Ortner M, Sippl M. Progress in fold recognition. Proteins 1995;23:376–386.
6. Lathrop RH, Smith TF. Global optimum protein threading with gapped alignment and empirical pair score functions. J Mol Biol 1996;255:641–665.
7. Rice DW, Eisenber D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
8. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;220:471–480.
9. Eisenberg D. Into the black of night. Nature Struct Biol 1997;4:95–97.
10. Finkelstein AV. Protein structure: what is it possible to predict now? Curr Opin Struct Biol 1997;7:60–71.
11. Marchler-Bauer A, Bryant SH. A measure of success in fold recognition. Trends Biochem Sci 1997;22:236–240.
12. Jones DT. Progress in protein structure prediction. Curr Opin Struct Biol 1997;7:377–387.
13. Levitt M. Competitive assessment of protein fold recognition and alignment accuracy. Proteins 1997;Suppl 1:92–104.
14. Rychlewski L, Zhang BH, Godzik A. Fold and function predictions for Mycoplasma genitalium proteins. Fold Des 1998;3:229–238.
15. Jaroszewski L, Rychlewski L, Zhang BH, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 1998;7:1431–1440.
16. Westhead DR, Thornton JM. Protein structure prediction. Curr Opin Biotech 1998;9:383–389.
17. Koehl P, Levitt M. A brighter future for protein structure prediction. Nature Struct Biol 1999;6:108–111.
18. Sternberg MJE, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 1999;9:368–373.
19. Jones DJ. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
20. Murzin AG. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. Proteins 1999;Suppl 3:88–103.
21. Jones DT, Tress M, Bryson K, Hadley C. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins 1999;Suppl 3:104–111.
22. Domingues FS, Koppensteiner WA, Jaritz M, Prlic A, Weichenberger C, Wiederstein M, Floeckner H, Lackner P, Sippl M. Sustained performance of knowledge-based potentials in fold recognition. Proteins 1999;Suppl 3:112–120.
23. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. Proteins 1999;Suppl 3:121–125.

24. Ota M, Kawabata T, Kinjo AR, Nishikawa K. Cooperative approach for the protein fold recognition. Proteins 1999;Suppl 3: 126–132.
25. Panchenko A, Marchler-Bauer A, Bryant SH. Threading with explicit models for evolutionary conservation of structure and sequence. Proteins 1999;Suppl 3:133–140.
26. Koretke KK, Russell RB, Copley RR, Lupas AN. Fold recognition using sequence and secondary structure information. Proteins 1999;Suppl 3:141–148.
27. Yang AS, Honig B. Sequence to structure alignment in comparative modeling using PrISM. Proteins 1999;Suppl 3:66–72.
28. Panchenko A, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
29. 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.
30. 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.
31. 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.
32. Muller A, MacCallum RM, Sternberg MJE. Benchmarking PSI-BLAST in genome annotation. J Mol Biol 1999;293:1257–1271.
33. Dunbrack RL Jr. Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. Proteins 1999;Suppl 3:81–87.
34. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234.
35. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
36. Fischer D, Eisenberg D. Fold recognition using sequence-derived properties. Protein Sci 1996;5:947–955.
37. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
38. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
39. Bryant SH. Evaluation of threading specificity and accuracy. Proteins 1996;26:172–185.
40. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. A method for the improvement of threading-based protein models. Proteins 1999;37: 592–610.
41. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein database. Science 1992;256:1433–1445.
42. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
43. Gerstein M, Hegyi H. Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev 1998;22:277–304.
44. Teichmann SA, Chothia C, Gerstein M. Advances in structural-genomics. Curr Opin Struct Biol 1999;9:390–399