

Single-Body Residue-Level Knowledge-Based Energy Score Combined With Sequence-Profile and Secondary Structure Information for Fold Recognition

Hongyi Zhou and Yaoqi Zhou

Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology & Biophysics, State University of New York at Buffalo, New York

ABSTRACT An elaborate knowledge-based energy function is designed for fold recognition. It is a residue-level single-body potential so that highly efficient dynamic programming method can be used for alignment optimization. It contains a backbone torsion term, a buried surface term, and a contact-energy term. The energy score combined with sequence profile and secondary structure information leads to an algorithm called SPARKS (Sequence, secondary structure Profiles and Residue-level Knowledge-based energy Score) for fold recognition. Compared with the popular PSI-BLAST, SPARKS is 21% more accurate in sequence-sequence alignment in ProSup benchmark and 10%, 25%, and 20% more sensitive in detecting the family, superfamily, fold similarities in the Lindahl benchmark, respectively. Moreover, it is one of the best methods for sensitivity (the number of correctly recognized proteins), alignment accuracy (based on the MaxSub score), and specificity (the average number of correctly recognized proteins whose scores are higher than the first false positives) in LiveBench 7 among more than twenty servers of non-consensus methods. The simple algorithm used in SPARKS has the potential for further improvement. This highly efficient method can be used for fold recognition on genomic scales. A web server is established for academic users on <http://theory.med.buffalo.edu>. *Proteins* 2004;55:1005–1013.

© 2004 Wiley-Liss, Inc.

Key words: knowledge-based potential; structure selections; fold recognition; threading; protein structure prediction

INTRODUCTION

Proteins adopt a limited number of unique structural folds. As a result, unrelated (or remotely related) sequences may encode similar structure. The aim of fold recognition is to recognize the structure similarities of proteins. Although significant advances have been made in the last 20 years, fold recognition continues to be one of the bottlenecks for accurate structure prediction in the CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments.¹

There are two main approaches for fold recognition. One approach is sequence-structure threading² in which the

compatibility of a sequence with each known structure is assessed by a score function (or structural profile).^{3–8} (For recent reviews, see References 9–13.) The other approach involves direct sequence-sequence comparison on the pairwise^{2,14–18} or multiple sequence^{19–25} levels. (For recent reviews, see References 12, 26–28.) Because the two approaches use different input information, a combined approach^{29–39} is superior than either approach alone. However, this combined approach often increases detection sensitivity but does not improve significantly the accuracy of alignment. Panchenko et al.³⁵ showed that improving both sensitivity and accuracy of alignment requires an optimal combination of structural and sequence profile terms.

In this paper, we develop a single-body knowledge-based energy score for fold recognition. A single-body energy score (sometimes called profile-energy score^{3,13}) allows the use of efficient dynamic programming for optimal alignment. This is in contrast to the pairwise interaction potential which would require extensive computational time^{40–42} or a frozen environment approximation⁴³ for optimal alignment. Previous single-body scores are limited to one or more of the properties such as hydrophobicity, solvent exposure, secondary structure, and contact energies. For example, Shan et al.³⁶ (COBLATH) and Kelley et al.³⁴ (3D-PSSM) used a score function that only takes the secondary structure and solvent exposure into account whereas Panchenko et al.³⁵ employed a contact potential. (For a review, see David et al.¹¹) In this paper, we establish a more elaborate knowledge-based score that contains a torsion-angle term for backbone interaction and a combined buried surface and contact-energy term for residue-residue and residue-solvent interactions. We found that a torsion-angle term is important for correctly identifying

Grant sponsor: National Institutes of Health; Grant numbers: R01 GM 066049 and R01 GM 068530M; Grant sponsor: Howard Hughes Medical Institute Center for Single Molecule Biophysics, State University of Buffalo; Grant sponsor: the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo.

*Correspondence to: Dr. Yaoqi Zhou, Howard Hughes Medical Institute Center for Single Molecule Biophysics and Department of Physiology & Biophysics,

State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214. E-mail: yqzhou@buffalo.edu

Received 9 May 2003; Accepted 5 September 2003

Published online 2 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20007

TABLE I. The Buriability, the Average Buried ASA, and Average C_α Contact Number of 20 Amino Acid Residues (cal/mol/Å²)

	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASNGLU	ASP	HIS	ARG	LYS	PRO	
B_k^{46}	22.6	15.7	23.9	20.3	20.8	19.5	24.5	19.5	13.4	7.0	10.3	8.2	8.5	7.6	7.3	8.2	11.3	8.5	6.1	9.9
\overline{dASA}^{48}	80	163	185	171	170	143	234	189	94	58	98	78	108	94	104	88	138	131	93	88
$\overline{N}_I^{\text{contact a}}$	13.5	12.1	12.2	12.7	12.3	12.7	12.0	12.0	12.0	11.4	11.3	11.1	10.4	10.6	9.7	10.0	11.3	10.7	9.9	10.6

^aFrom a database of 1011-protein database.⁴⁵

the native structure from decoys and this term together with a buried surface and contact energy terms provides one of the best residue-level potentials for structure selections from decoys.

The proposed knowledge-based, structure-derived score function is combined with the sequence profiles generated from PSI-BLAST²¹ and the secondary structure information predicted from PSIPRED⁴⁴ for fold recognition. The highly efficient method called SPARKS (Sequence, secondary-structure Profiles And Residue-level Knowledge-based Score) is found to improve over most existing, nonconsensus, fold-recognition methods in sensitivity, alignment accuracy, and specificity based on several well-established benchmarks.

METHODS

The Knowledge-Based Score

The knowledge-based, structure-derived score function for residue type I at the position i is a sum of three terms (a backbone torsion term [$u^{\text{torsion}}(I, \phi_i, \psi_i)$], a buried surface term [$u^{\text{surface}}(I, dASA_i)$], and a contact energy term [$u^{\text{contact}}(I, N_i^{\text{contact}})$]):

$$u_i^{\text{structure}}(I) = u^{\text{torsion}}(I, \phi_i, \psi_i) + w^{\text{surface}} u^{\text{surface}}(I, dASA_i) + w^{\text{contact}} u^{\text{contact}}(I, N_i^{\text{contact}}), \quad (1)$$

where w^{surface} and w^{contact} are weight factors for buried surface and contact energy terms relative to the torsion term, respectively, $dASA_i$ and N_i^{contact} are the buried accessible surface area (ASA) and the number of contacts for the C_α atom of the residue type I at position i , respectively, and $\phi_i = \phi(C_{i-1}-N_i-C_{\alpha i}-C_i)$ and $\psi_i = \psi(N_i-C_{\alpha i}-C_i-N_{i+1})$ are the backbone torsion angles at each side of a residue at the sequence position i . [For the first and last residues, $u^{\text{torsion}}(I, \phi_i, \psi_i)$ is set to zero because torsion angles are not defined.]

The torsion potential $u^{\text{torsion}}(I, \phi_i, \psi_i)$ is given by

$$u^{\text{torsion}}(I, \phi_i, \psi_i) = -RT \ln \frac{N_{\text{obs}}(I, \phi_i, \psi_i)}{\sum_{\phi_i, \psi_i} N_{\text{obs}}(I, \phi_i, \psi_i) / N_{\text{bin}}^2}. \quad (2)$$

In this equation, the torsion potential is the logarithm of the number of observed occurrence of the residue type I at torsion angles of ϕ_i and ψ_i [$N_{\text{obs}}(I, \phi_i, \psi_i)$] normalized by the averaged occurrence. In the calculations, torsional angles ϕ_i and ψ_i at each site i are divided into 36 bins. That is, $N_{\text{bin}} = 36$. The potential parameters contained in a table of $20 \times 36 \times 36$ are available on <http://theory.med.buffalo.edu>. The statistical distribution is obtained by

using a structural database of 1011 non-homologous soluble proteins.⁴⁵

The buried accessible surface term of residue type I is given by

$$u^{\text{surface}}(I, dASA_i) = -B_k^I \times dASA_i, \quad (3)$$

where B_k^I is the buriability of residue type I and $dASA_i$ is the buried accessible surface area at the position i . This term is introduced because it was shown that the average contribution of a residue to the stability of a protein is linearly correlated with its buried accessible surface area with a slope called buriability.⁴⁶ The value of buriability was derived from an all-atom distance-dependent potential of mean force using a distance-scaled infinite-ideal gas reference state (DFIRE)⁴⁷ and the structural database of 1011 proteins.⁴⁵ The value of buriability for each amino acid,⁴⁶ listed in Table I, can quantitatively take into account the fact that hydrophobic residues are more likely to be buried into the core of proteins.⁴⁶

To increase the computational efficiency, we take advantage of the fact that the buried ASA ($dASA$) normalized by the average buried ASA (\overline{dASA}) correlates strongly with the number of contacts (N^{contact}) normalized by the average number of contacts ($\overline{N^{\text{contact}}}$). As shown in Figure 1, the correlation is not perfect but significant enough for the purpose of this paper. That is, Eq. (3) can be rewritten as

$$u^{\text{surface}}(I, dASA_i) \approx u^{\text{surface}}(I, N_i) = -B_k^I \times \frac{\overline{dASA}^I}{\overline{N}_I^{\text{contact}}} \times N_i^{\text{contact}}, \quad (4)$$

where \overline{dASA}^I and $\overline{N}_I^{\text{contact}}$ are average buried ASA obtained from Zhou and Zhou⁴⁸ and average C_α contact number of a residue type calculated from the 1011-protein database,⁴⁵ respectively. The values of \overline{dASA}^I and $\overline{N}_I^{\text{contact}}$ are also listed in Table I.

The contact energy term is given by

$$u^{\text{contact}}(I, N_i^{\text{contact}}) = -RT \ln \frac{N_{\text{obs}}^{\text{contact}}(I, k)}{\sum_k N_{\text{obs}}^{\text{contact}}(I, k) / N_{\text{cbin}}} \quad (5)$$

where $N_{\text{obs}}^{\text{contact}}(I, k)$ is the number of observed contacts of residue I with other residues at k 'th bin and N_{cbin} , the number of contact bins, is set to 25. In other words, the number of contacts allowed in the program is 25. In the rare occasions of more than 25 contacts, the statistics is included in the bin for 25 contacts. Here, we set $T = 300K$ ($RT = 0.6$ kcal/mol). When $N_{\text{obs}}^{\text{contact}}(I, k) = 0$, the knowledge-

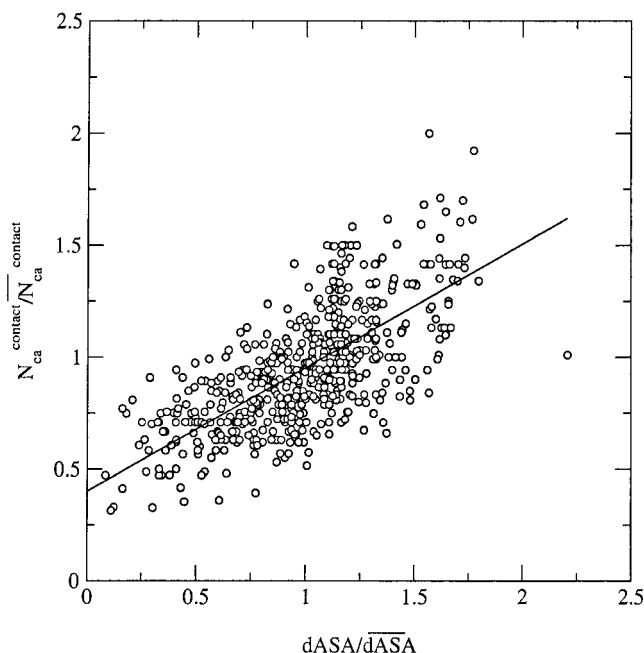


Fig. 1. The buried ASA ($dASA$) normalized by the average buried ASA (\overline{dASA}) correlates strongly with the number of contacts (N_{contact}) normalized by the average number of contacts ($\overline{N_{\text{contact}}}$) based on the distance between C_{α} atoms. The correlation coefficient is 0.71 for 1023 data points.⁴⁸

based score $u_i^{\text{structure}}(I)$ is set to a constant of 10 kcal/mol. A contact is defined by the C_{α} – C_{α} distance within 8.5 Å. The 20×25 matrix for $u_{\text{contact}}^{\text{structure}}(I, N_i^{\text{contact}})$ is available in the SPARKS web server.

It should be noted that buried surface and contact-energy terms are two different ways of calculating the potential of mean force for residue-residue and residue-solvent interactions (See, e.g. Melo et al.⁴⁹). One is expressed in term of buried surface area and is proportional to the number of contacts whereas the other is proportional to the logarithm of the number of contacts. Because it is not clear about which expression is a more accurate representation for the potential of mean force, we include both terms by introducing adjustable weight factors [Eq. (1)]. These weight factors are to be determined by optimizing the performance for selecting structures from decoys (see below). The total energy score of a protein is given by

$$u_{\text{total}}^{\text{structure}} = \sum_{i=2}^{N_r-1} u_i^{\text{structure}}(I), \quad (6)$$

where N_r is the total number of residues, I is the residue type of residue at position i . Note that the first and last residues are excluded from the calculations. Using the decoy sets of *4state_reduced*,⁵⁰ *fisa*,⁵¹ *fisa casp3*,⁵¹ *lmds*, and *lattice ssfit*,⁵² we found that optimal values of w_{contact} and w_{surface} are 1.0 and 2.0, respectively. These parameters lead to an energy score that correctly identified 22 native proteins from a 32 protein decoy sets. This success rate was similar or better than some distance-dependent

all-atom pair potentials⁴⁷ (RAPDF⁵³ and atomic KBP⁵⁴), which are more accurate than the corresponding residue-level potentials. Furthermore, the best distance-dependent potential (TE-13) among several residue-level potentials compared by Tobi and Elber⁵⁵ identified 14 native structures out of 25 decoy sets. The success rate of the new score function [Eq. (6)] is 17 out of the same 25 proteins. Thus, the new single-body, distance-independent, potential is one of the best residue-level potentials for structure discrimination.

Sequence Profiles

The log-likelihood (position specific score matrix PSSM) profile $P_j^b(t)$ of a sequence b of residue type t at position j was generated by PSI-BLAST²¹ after three iterations and an E value cutoff of 0.001. This is the same setting used by PSIPRED for secondary structure prediction.⁴⁴ The sequence database was obtained from NCBI at ftp://ftp.ncbi.nih.gov/blast/db/nr.Z dated November 22, 2002.

A frequency profile F_i^{query} of the query sequence at position i was also constructed from the output of the multiple sequence alignment of PSI-BLAST for those sequences that have less than 98% identity and less than 0.001 E-value. The profile was normalized to one at each sequence position. The mutation score is expressed by profiles as follows:

$$u^{\text{mutation}}(i, j) = - \sum_{t=1}^{20} F_i^{\text{query}}(t) \cdot P_j^{\text{template}}(t), \quad (7)$$

Where F_i^{query} [P_j^{template}] is the frequency (log-likelihood) profile of query sequence (template sequence), and the summation is over the 20 residue types. This mutation score was also used by others for fold recognition (e.g., Fischer⁵⁶).

Secondary Structures

The log-likelihood profile of query sequence, $P_j^{\text{query}}(t)$, was used for the secondary structure prediction by PSIPRED. The score from the predicted secondary structure information is given by

$$u^{\text{2ndary}}(i, j) = -\delta_{s^{\text{query}}(i), s^{\text{template}}(j)}, \quad (8)$$

where $s^{\text{query}}(i)$ and $s^{\text{template}}(j)$ are the predicted secondary structure of the query sequence at position i and the actual secondary structure of the template at position j , respectively and $\delta_{s^{\text{query}}(i), s^{\text{template}}(j)}$ is 1 if there is a match between predicted and actual secondary structures and -1 , if otherwise. Similar score function for secondary structure was used by Fischer and Eisenberg.³¹ The secondary structures of templates were obtained by H-bonds (DSSP-like) criteria.⁵⁷ Three states (helix, strand, coil) were used for secondary structure. A built-in simplified PSIPRED algorithm of secondary structure prediction is used to speed up computation. The original PSIPRED used two rounds of neural networks and provided a confidence level (C. L.) for the prediction at each position. Our implementation employed only one round of neural network without evaluation of C. L. This is because we found that a more

accurate prediction of the secondary structure by the second neural network did not further improve the accuracy of fold recognition.

The Fold Recognition Algorithm and Assessment

The sequence profile and secondary structure information is combined with the knowledge-based score for fold recognition. The total score is given by

$$u^{\text{total}} = u^{\text{mutation}}(i, j) + w^{\text{2ndary}} u^{\text{2ndary}}(i, j) + w^{\text{structure}} u_j^{\text{structure}}(t_i), \quad (9)$$

where $u_j^{\text{structure}}(t_i)$ is the knowledge-based score of the template at position j with the residue type t_i of the query position i , and w^{2ndary} and $w^{\text{structure}}$ are the weight factors for secondary structure and structure scores, respectively. As with Fischer and Eisenberg,³¹ we let w^{2ndary} equal to the gap extension penalty w_1 to reduce the number of adjustable parameters.

A profile-profile alignment method using a global-local dynamic programming algorithm⁵⁸ was employed to find the minimum of the total score that aligns the query sequence with a template in the template library. That is, the raw score of an alignment is u^{total} . The Z_{score} for a given template is defined as $(u^{\text{total}}_{\text{template}} - u_{\text{ave}})/u_{\text{sd}}$, where u_{ave} and u_{sd} are the average and the standard deviation of the raw scores for all templates, respectively. The initial ranking is given by the raw score normalized by the alignment length (S_{template}) in order to remove sequence-length dependence. To determine the specificity of the fold recognition method, the significance score of a template is calculated by the multiplication of its normalized raw score (S^{template}) and the Z_{score} . A bonus score of -1 is also added to the significance score if the model has a structural similarity with one or more of the next nine models based on initial ranking with normalized raw scores. (A structure is similar if MaxSub score is greater than 0.01). This simple method for calculating significance of an alignment is based on the rationale that a good model should not only have a very low normalized score (S^{template}) but also have a score that is lower relative to the scores based on other templates (Z_{score}). A similar method is used in Ref. 36.

All weight factors and gap parameters were obtained by optimizing the performance of the method in Fischer's dataset³¹ (see Results). $w_0 = 4.5$, $w_1 = 0.5$, $w^{\text{2ndary}} = w_1$, and $w^{\text{structure}} = 0.175$ whereas $w^{\text{contact}} = 1.0$ and $w^{\text{surface}} = 2.0$ by optimizing native-structure selections from decoys (see above).

The accuracy of alignment was assessed by MaxSub score⁵⁹ which is a measure of similarity between 0.0 (no similarity) and 1.0 (perfect similarity) between the predicted (model) structure and the native structure. The value is calculated by searching the largest subset of well-superimposed residues (≤ 3.5 Å). The model structure was built by using the first-ranking template with MODELLER 6 v1.⁶⁰ We use MaxSub score because it is the official evaluation method used in the CAFASP (Critical assessment of fully automated protein structure predic-

tion methods) experiments (<http://www.cs.bgu.ac.il/dfischer/cafasp1/cafasp1.html>).

RESULTS

Training Set: Fischer's Dataset for Sensitivity and Alignment Accuracy

The initial gap penalty (w_0) and the extension gap penalty (w_1) of the SPARKS method were first optimized by using Fischer's dataset³¹ without the knowledge-based, structure-derived score. The dataset contains 68 probe sequences and 301 library structures. A match occurs when the expected match was ranked as the number one or only below its superfamily members based on SCOP 1.61 classification (i.e., no incorrect folds have a better score than the expected match⁶¹). The highest success rate is 56/68 when $w_0 = 4.5$ and $w_1 = 0.5$. The weight factor for the secondary structure w^{2ndary} is set to be same as w_1 (see Methods).

The weight factor for the structure-derived knowledge-based score $w^{\text{structure}}$ was then obtained by including the single-body knowledge-based score and further optimizing the performance for Fischer's dataset. We found that the number of correct matches is not very sensitive to the value of $w^{\text{structure}}$. Thus, we optimized $w^{\text{structure}}$ for alignment accuracy (MaxSub score⁵⁹) instead. The total MaxSub score for Fischer's dataset was maximized to 26.12 at $w^{\text{structure}} = 0.175$. This score increases from 24.68 given by SPARKS without the structure-derived score. The number of correctly recognized proteins for SPARKS, however, remains 56. This success rate (56/68) is the same as that of COBLATH (a combined approach of sequence profile with a score matrix for secondary structure and solvent exposure)³⁶ but is lower than that of a computationally more intensive, hierarchical threading method called PROSPECTOR (58-61/68),³⁹ which involves pair interaction potentials and separate profiles for closely-related and distant-related sequences.

Test Set 1: ProSup Benchmark for Sequence-Alignment Accuracy

ProSup benchmark was prepared by Sippl's group⁶² to test the alignment accuracy of fold recognition methods. The set consists of 127 pairs of proteins with correct alignments obtained by structural alignment program ProSup. The accuracy of an alignment was obtained by calculating the percent of matches between the correct alignment and the alignment made by a fold recognition method. Table II compares the performance of several methods. It should be emphasized, however, that the performance of our method is not directly comparable with the published results for some of the previous methods because available sequence and structure databases at the time is different. Nevertheless, the comparison can be served as an approximate indicator for the accuracy of SPARKS. For the direct pairwise sequence alignment, the optimized substitution matrix STROMA¹⁸ performs the best. It is remarkable that an optimized substitution matrix can perform even slightly better than multiple-sequence alignment method of PSI-BLAST.²¹ A threading

TABLE II. The Alignment Accuracy for ProSup Benchmark

Method	FASTA ^a	Sequence ^a	Stroma ^b	PSI-BLAST	Threading ^c	SPARKS ^d
Accuracy	31.4%	34.1%	36.1%	35.6%	48.0%	57.2%(51.4%)

^aPairwise sequence comparisons. Results from Sippl et al. (Domingues et al.⁶²).

^bPairwise sequence comparison using the optimized substitution matrix STROMA from Quian and Goldstein¹⁸

^cResults from Domingues et al.⁶²

^dThis work, the number in parenthesis is the result of SPARKS without the structure-derived score.

TABLE III. Performance of Different Methods for Fold Recognition in Lindahl Benchmark

Method	Family only		Superfamily only		Fold only	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
THREADER ^a	49.2%	58.9%	10.8%	24.7%	14.6%	37.7%
HMMER-PSIBLAST ^a	67.7%	73.5%	20.7%	31.3%	4.4%	14.6%
SAMT98-PSIBLAST ^a	70.1%	75.4%	28.3%	38.9%	3.4%	18.7%
BLASTLINK ^a	74.6%	78.9%	29.3%	40.6%	6.9%	16.5%
SSEARCH ^a	68.6%	75.5%	20.7%	32.5%	5.6%	15.6%
PSI-BLAST ^a	71.2%	72.3%	27.4%	27.9%	4.0%	4.7%
FUGUE ^a	82.2%	85.8%	41.9%	53.2%	12.5%	26.8%
SPARKS ^b	81.6%(82.9%)	88.1%(90.1%)	52.5%(56.5%)	69.1%(72.4%)	24.3%(23.1%)	47.7%(43.6%)

^aFrom Shi et al.⁶⁴ The upgrade versions of these methods with current sequence and structure databases may perform better than the above results.

^bThe numbers in parenthesis are the results of SPARKS without the structure-derived score.

method given by Sippl et al.⁶² makes a significant improvement ($> 10\%$) over sequence-based methods. However, the sequence profile and secondary structure information in SPARKS alone is already more accurate than the threading method (51.4% versus 48.0%). Incorporation of the single-body knowledge-based score increases the accuracy by another 5.3% to 57.2%. The latter is more than 21% better than the accuracy provided by PSI-BLAST. If the criterion for the accuracy is relaxed to allow a shift of ≤ 4 residues from the “correct” alignment, the overall accuracy of alignment is 78.5% for SPARKS (69.2% for the SPARKS without the knowledge-based score). Note that the improvement upon the use of the structure-derived score is 8.9% for a relaxed criterion for the accuracy. This is a significant increase from 5.3% based on one-to-one match. Clearly, introducing the knowledge-based score brings in the significant improvement in alignment accuracy.

Test Set 2: Lindahl Benchmark for Fold-Recognition Sensitivity

The Lindahl set⁶³ was designed to assess the fold recognition sensitivity. It has 976 proteins. Each protein is aligned with the rest 975 proteins. There are 555, 434, and 321 pairs of proteins in the same family, superfamily, and fold, respectively. The fold-recognition method is tested by checking whether or not the method can recognize the member of same family, superfamily, or fold as the first rank or within the top five ranks. The results of SPARKS are compared with several well-established methods in Table III. We stress that the comparison only serves an approximate guide because the sequence and structure database available for previous methods are smaller than the one used in SPARKS. It shows that SPARKS (with or without knowledge-based scores) is the most sensitive method to detect structural similarity among all the

methods listed in Shi et al.⁶⁴ (except at the family level where the performance of FUGUE is comparable to that of SPARKS). For example, compared to the popular PSI-BLAST, SPARKS is 10%, 25%, and 20% more sensitive in recognizing the member of same family, superfamily, and fold, respectively. However, all sensitivities except at the fold level seem to decrease somewhat after the inclusion of the single-body knowledge-based score in SPARKS (Table II). This sensitivity test, however, may not reflect the true alignment accuracy because the result is based on somewhat subjective SCOP classification.⁶⁵ To address this question more quantitatively, we calculated the MaxSub score between the model built from first-ranking template and the known native structure. It is found that the total number of recognized proteins (MaxSub > 0.01) increases from 616 to 630 (a 2.0% increment) whereas the total MaxSub score increases from 322.72 to 334.47 (a 3.6% increment). Thus, the knowledge-based score makes a small but significant improvement in both sensitivity and accuracy of fold recognition for the Lindahl benchmark set.

Test Set 3: LiveBench 7

Sixty proteins of LiveBench 7⁶⁶ targets released between February 1, 2003 and April 10, 2003 are used to further test the performance of SPARKS (<http://bioinfo.pl/LiveBench/>). This partial LiveBench set is what was available prior to the preparation and final submission of this paper. The LiveBench targets were new entries in the Protein Data Bank which do not have sequence similarity to previously released proteins and thus provide an unbiased set for testing the performance of fold recognition methods. Currently, more than 30 servers are tested on LiveBench automatically.

The template library for SPARKS was built by using the 40% representative domains of SCOP 1.61. The entire

TABLE IV. Performance for the 60 LiveBench 7 Targets[†]

Method ^b	Sensitivity ^c	Overall accuracy ^d	Specificity ^a					Average
			1	2	3	4	5	
3D-PSSM ³⁴	39	13.36	19	31	32	32	36	30.0
PROSPECT II ³⁸	41	13.52	33	35	35	35	36	34.8
FFAS03 ²³	41	13.98	32	34	36	38	38	35.6
SPARKS ^e	46(40)	14.40(13.68)	31(26)	40(30)	40(34)	41(34)	41(35)	38.4(31.8)
3DS5 ^f	49	17.19	39	42	43	46	47	43.4

[†]List of 60 proteins: 1gvnb, 1ghea, 1nb8a, 1ns5a, 1nlxa, 1n9pa, 1n91a, 1lj9a, 1mlga, 1m33a, 1mzga, 1mwka, 1mx0a, 1n3ja, 1nbwb, 1lq9a, 1ni5a, 1l7aa, 1n3ka, 1nkva, 1kyta, 1gz5a, 1nh7a, 1nnwa, 1iyga, 1o70a, 1iq6a, 1nekd, 1nekc, 1n05a, 1j85a, 1l3pa, 1n81a, 1n67a, 1on0a, 1omia, 1olya, 1nxza, 1nxua, 1lzoa, 1hkea, 1nyla, 1n57a, 1ngna, 1m5qa, 1kkga, 1gpqa, 1o7de, 1l9ka, 1nrja, 1nafa, 1n2ma, 1gwka, 1o22a, 1o20a, 1olza, 1olxa, 1nt2b, 1izna, 1liznb

^aThe specificity is defined as the numbers of recognized proteins have scores better than 1–5 false positives.

^bResults of other methods are taken from the top three performers of non-consensus methods in LiveBench 7 (not including the single server with built-in multiple methods). Their MaxSub scores are calculated locally based on the predicted models.

^cSensitivity is the number of targets whose first-ranking models with a MaxSub score of greater than 0.01.

^dTotal MaxSub score.

^eThe numbers in parenthesis are the results of SPARKS without the structure-derived score.

^f3D-SHOTGUN consensus method.⁶⁹

chains of multiple-domain proteins are contained in the library. The library was then updated with new proteins released before February 3, 2003 if they have less than 40% sequence identity with the sequences already in the library. (This was done by protein sequence culling server PISCES⁶⁷).

The results of SPARKS were compared with the then top three single-method performers of more than 20 servers of non-consensus methods in Table IV. The number of recognized proteins (sensitivity) increases from between 39 and 41 for the top three performers (a sequence-based method: FFAS03²³ and two combined methods: 3D-PSSM³⁴ and PROSPECT II³⁸) to 46 for SPARKS. SPARKS also provides the most accurate model structures based on the total MaxSub score and has the highest specificity. The total MaxSub score increases from between 13.36 and 13.98 for the top three performers to 14.40 for SPARKS and the specificity increases from between 30.0 and 35.6 to 38.4. In Table III, the effect of the knowledge-based score is also illustrated. The performance of SPARKS without the structure-derived score is comparable to the other top three performers in sensitivity and overall alignment accuracy but a slightly lower value for specificity. The use of the structure-derived score leads to a marked improvement in all three categories (15%, 5%, and 21% in sensitivity, model accuracy, and specificity, respectively). It should be noted that the ranking score based on $S^{\text{template}}Z_{\text{score}}$ (see Methods) is perhaps the simplest score function for determining specificity among the top performers. For example, FFAS03 uses a sophisticated transformation of Z_{score} . The 3D-PSSM calculates a theoretical error rate per query (ET).

The above results are limited to 60 targets. The performance of SPARKS for the entire 115 targets on LiveBench 7⁶⁶ (<http://bioinfo.pl/LiveBench/>) is available at the resubmission. Table V compares the results of SPARKS with those of the top performers of single servers. All results were directly taken from LiveBench server. SPARKS provides one of the most accurate model structures based on the total MaxSub score and has one of the highest

TABLE V. Performance for the 115 LiveBench 7 Targets[†]

Method	Sensitivity ^a	Total MaxSub	
		score	Specificity ^b
INBGU ^{56c}	65	22.34	56.3
SHGU ^d	69	25.56	57.4
SHUM ^d	71	25.27	55.1
FFAS03 ²³	70	24.78	63.3
RAPT ⁶⁸	69	24.31	52.2
FUG2 ⁶⁴	63	21.90	57.9
SPARKS	65	23.85	55.8
3DS5 ^e	79	29.96	68.3

[†]Taken from LiveBench server. Listed are individual servers having either better MaxSub score or specificity or both than SPARKS. Meta server 3DS5 is included for comparison.

^aSensitivity is the number of targets whose first-ranking models with a MaxSub score of greater than 0.01.

^bThe specificity is define as the average number of recognized proteins that have scores better than 1–10 false positives.

^cINBGU is a combination of five methods which exploit sequence and structure information in different ways and produces one consensus prediction of the five.

^dShotGun-INBGU uses the ShotGun method to create alternative consensus models for the INBGU components. SHUM is an experimental version.

^e3D-SHOTGUN meta server.⁶⁹

specificities. The total MaxSub score is 23.85 behind SHGU(25.56), SHUM (25.27), FFAS03 (24.78), and RAPTOR (24.31). The specificity of SPARKS (55.8) is also ranked #5 behind FFAS03 (63.3), FUG2 (57.9), SHGU (57.4) and INBGU (56.3). The number of recognized proteins (sensitivity) is 65. This is rank #7 behind SHUM (71), FFAS03²³ (70), RAPTOR⁶⁸ (69), SHGU (69), 3DPS (67) and PRO2 (66). (The results for the last two methods are not shown.) Among the top performers on the LiveBench server, INBGU, SHGU and SHUM are single servers with built-in multiple methods. INBGU⁵⁶ is a combination of five methods which exploit sequence and structure information in different methods and produces one consensus prediction of the five. SHGU(ShotGun-INBGU) uses the ShotGun method to create alternative consensus models

TABLE VI. Reevaluation of the Performance for the 115 LiveBench 7 Targets

Method	Sensitivity ^a	Total MaxSub score	Specificity ^b
RAPT ⁶⁸	74	24.84	54.8
FFAS03 ²³	80	26.22	66.7
SHGU ^c	88	28.15	67.3
SPARKS	82(80) ^d	25.76(24.98)	71.6(59.1)

^aSensitivity is the number of targets whose first-ranking models with a MaxSub score of greater than 0.01.

^bThe specificity is defined as the average number of recognized proteins that have scores better than 1–10 false positives.

^cShotGun-INBGU uses the ShotGun method to create alternative consensus models for the INBGU components.

^dThe numbers in parenthesis are the results of SPARKS without the structure-derived score.

for the INBGU components. SHUM is an experimental version. FFAS is a sequence-based method. RAPTOR, FUG2, 3DPS, and PRO2, similar to SPARKS, use a combined sequence and structural information.

It should be noted that all single methods are not as accurate as some consensus methods. The results of the best performer 3D-SHOTGUN⁶⁹ were also shown in Tables III and IV for comparison.

A careful review of our results found that MaxSub scores for some models are reported as zero on the LiveBench server but are nonzero when calculated locally. (We are able to reproduce all nonzero values of MaxSub scores.) This happens not only to SPARKS but also to FFAS03 or other methods. For example, the MaxSub scores for the target 1kkga are zero for SPARKS and FFAS03 on the server but are 0.177 (template 1fxka) for SPARKS and 0.195 (template 1jmta) for FFAS03, respectively, when calculated locally. We found that the difference is mostly caused by the way to handle missing structural regions in the native structures and nonstandard residues. Because all query sequences are treated as continuous, the corresponding pdb files are changed accordingly. Table VI compares the results of sensitivity, accuracy, and specificity that are reevaluated for FFAS03, RAPTOR, SHGU, and SPARKS. (We only picked a few from Table IV that are mostly likely to perform better than SPARKS.) For FFAS03, RAPTOR and SHGU, we downloaded all the models predicted by them from the LiveBench server. As the table shows, the results for all four methods improve over the published values on the server. SPARKS now has a higher sensitivity (82) and specificity (71.6) than FFAS03 (80 and 66.7, respectively) but have a lower accuracy (25.76) than FFAS03 (26.22) in term of total MaxSub scores. The performance of SPARKS are better than that of RAPTOR in all categories. Compared to SHGU, SPARKS continues to have the highest specificity but has lower sensitivity and accuracy. Thus, the reevaluation of the results significantly improves the rank of SPARKS relative to others.

DISCUSSION

In this paper, we have designed a single-body knowledge-based score for fold recognition. The score function con-

tains a backbone torsion term, a buried surface term, and a contact-energy term. Each term is a statistical potential derived from 1011 non-homologous structure database of proteins.⁴⁵ The first and last two terms are designed to take the backbone and side-chain interactions into account, respectively. The relative weights of the three terms were optimized by using well-established 32 decoy sets (Methods). The structure-derived score function is then combined with the sequence profile from PSI-BLAST and the secondary structure information from PSIPRED to obtain a new method for fold recognition (SPARKS). SPARKS is shown to improve sensitivity, alignment accuracy, and specificity over many well-established non-consensus methods. The improvement over other methods is significant in all testing benchmarks.

The success of SPARKS illustrated the power of combining the sequence-based and threading-based approaches in fold recognition.^{29–39} Much of the success is owing to the fact that the sequence profile and the secondary structure information alone yielded a method that has comparable accuracy with many existing well-developed algorithms. Another possible reason is that the SPARKS method, for the first time, incorporates sophisticated backbone and side interactions into the score function. We found that the lack of torsion angle term would significantly reduce the ability of the residue-based potential to discriminate against decoys. Moreover, there is a buried-surface term which quantitatively takes into account the fact that hydrophobic residues are more likely to be buried.⁴⁶

To further analyze the relative contributions of different terms in the structural score, we test our potential in all other six combinations (torsion only, ASA only, contact only, ASA + contact, torsion + contact, ASA + torsion) without adjusting the weight factors. The results on Prosop benchmark are shown in Table VII. The contact term does not seem to help for alignment accuracy. There is no change in alignment accuracy with the addition of contact term (51.4% versus 51.5%). However, when it was combined with ASA term, the two-term structural score (ASA + contact) becomes the second best (56.8%) in alignment accuracy behind SPARKS (57.2%). The ASA term appears to be the most important. It has the highest accuracy in single-term structure-score function and its combination with either contact or torsion terms yields the best and second best alignment accuracy in the two-term structure-score function. SPARKS (all three terms) provides a small but nontrivial improvement. If the criterion for the accuracy is relaxed to allow a shift of ≤ 4 residues from the “correct” alignment, the overall picture remains the same, i.e., the ASA term is the most important in the structural score. SPARKS is perhaps one of the simplest methods among the top performers. INBGU, SHGU, and SHUM are single servers with built-in multiple methods. FFAS03 uses a sophisticated transformation of Z_{score} to rank models. RAPTOR uses pair (two-body) interactions used in PROSPECT II and thus is computationally more expensive. Thus SPARKS has the potential for further improvement in accuracy of folding recognition.

TABLE VII. The Contribution to Alignment Accuracy for ProSup Benchmark by Various Terms in the Structure-score Function

Method	No structure score	Torsion	ASA	Contact	ASA + contact	Torsion + contact	ASA + torsion	SPARKS
Accuracy ^a	51.4%	53.0%	55.8%	51.5%	56.8%	53.5%	56.4%	57.2%
± 4 Residues ^b	69.2%	72.2%	77.5%	70.6%	77.8%	73.0%	78.1%	78.5%

^aAccuracy defined by one-to-one match given by the method and by the benchmark.

^bAccuracy defined by the match within four residues from the one-to-one match.

The use of a single-body potential in SPARKS leads to an efficient algorithm for fold recognition on a genomic scale. We find that the completion of fold recognition of *M. genitalium* genome (480 genes or proteins, <ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/>) takes only about a week on a PC with an Athlon 1.4GHz processor without serious effort on code optimization. Thus, an analysis of human genome (30,000–40,000 genes) using SPARKS can be completed within a few weeks using an inexpensive, small cluster of PCs. Thus, SPARKS can serve as a powerful new tool in structural genomics. Fold recognition is increasingly important as more and more structures with new folds become available. For example, we can resolve 71% of proteins in *M. genitalium* genome based on a cutoff of -2.9 for specificity ranking score that allows zero false positive in LiveBench 7 (Table VI).

ACKNOWLEDGMENT

We thank Professor D. Fischer for providing us his benchmark and Drs. Y. Xu and D. Xu for their help for installing and using their PROSPECT II program.

REFERENCES

- Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins* 2001;Suppl 5:2–7.
- Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature* 1992; 358:86–89.
- 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.
- 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.
- Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 1993;16:92–112.
- Abagyan R, Frishman D, Argos P. Recognition of distantly related proteins through energy calculations. *Proteins* 1994;19:132–140.
- Murzin AG, Bateman A. Distance homology recognition using structural classification of proteins. *Proteins* 1997;Suppl 1:105–112.
- Xu Y, Xu D. Protein threading using PROSPECT: Design and evaluation. *Proteins* 2000;40:343–354.
- Jones DT. Progress in protein structure prediction. *Curr Opin Struct Biol* 1997;7:377–387.
- Torda AE. Perspectives in protein-fold recognition. *Curr Opin Struct Biol* 1997; 7:200–205.
- David R, Korenberg MJ, Hunter IW. 3D-1D threading methods for protein fold recognition. *Pharmacogenomics* 2000;1:445–455.
- Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. *Proteins* 2001;Suppl 5:55–67.
- Meller J, Elber R. Protein recognition by sequence-to-structure fitness: Bridging efficiency and capacity of threading models. *Adv Chem Phys* 2002;120:77–130.
- Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. *Methods Enzymol* 1983;91:524–545.
- Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. *Proc Natl Acad Sci USA* 1988;85:2444–2448.
- Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment tool. *J Mol Biol* 1990; 215:403–410.
- Vingron M, Waterman MS. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J Mol Biol* 1994;235:1–12.
- Qian B, Goldstein RA. Optimization of a new score function for the generation of accurate alignments. *Proteins* 2002;48:605–610.
- Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 1998;14:846–856.
- Henikoff S, Henikoff JG. Amino acid substitutes matrices from protein blocks. *Proc Natl Acad Sci* 1992;89:10915–10919.
- 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.
- Bailey TL, Gribskov M. Score distributions for simultaneous matching to multiple motifs. *J Comput Biol* 1997;4:45–59.
- 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.
- Koretke KK, Russell RB, Lupas AN. Fold recognition from sequence comparisons. *Proteins* 2001;Suppl 5:68–75.
- Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol* 2002;315:1257–1275.
- Eddy S.R. Profile hidden Markov models. *Bioinformatics* 1998;14: 755–763.
- Henikoff S, Henikoff JG. Amino acid substitutes matrices. *Adv Protein Chem* 2000;54:73–96.
- Koonin EV, Wolf YI, Aravind L. Protein fold recognition using sequence profiles and its application in structural genomics. *Adv Protein Chem* 2000;54:245–275.
- Yi TM, Lander ES. Recognition of related proteins by iterative template refinement (ITR). *Protein Sci* 1994;3:1315–1328.
- Elofsson A, Fischer D, Rice DW, Le Grand SM, Eisenberg D. A study of combined structure/sequence profiles. *Fold Des* 1996;1: 451–461.
- Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. *Protein Sci* 1996;5:947–955.
- Rost B, Sander C. Protein fold recognition by prediction-based threading. *J Mol Biol* 1997;270:471–480.
- Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci* 1998;7:1431–1440.
- Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 2000; 299:499–520.
- Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. *J Mol Biol* 2000;296:1319–1331.
- Shan YB, Wang GL, Zhou HX. Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. *Proteins* 2001;42:23–37.
- Al-Lazikani B, Sheinerman FB, Honig B. Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc Natl Acad Sci USA* 2001;98:14796–14801.

38. Kim D, Xu D, Guo J, Ellrott K, Xu Y. PROSPECT II: Protein structure prediction program for the genome-scale. *Prot Eng* 2003;16:651–657.
39. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins* 2001;42:319–331.
40. Lathrop RH, Smith TF. Global optimum protein threading with gapped alignment and empirical pair score functions. *J Mol Biol* 1996;255:641–665.
41. Lathrop RH. The protein threading problem with sequence amino-acid interaction preferences is NP-complete. *Protein Eng* 1994;7:1059–1068.
42. De La Cruz X, Thornton JM. Factors limiting the performance of prediction-based fold recognition methods. *Protein Sci* 1999;40:750–759.
43. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. The statistical mechanical basis of sequence alignment algorithms for protein structure prediction. In: Elber R., editor. *Recent developments in theoretical studies of proteins*. Singapore: World Scientific 1996; 359–388.
44. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
45. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. *Protein Sci* 1992;1:409–417.
46. Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* 2004;54:315–322.
47. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 2002;11:2714–2726. Corrections. *Protein Sci* 2003;12:2121.
48. Zhou H, Zhou Y. The stability scale and atomic solvation parameters extracted from 1023 mutation experiments. *Proteins* 2002;49:483–492.
49. Melo F, Sanchez R, Šali A. Statistical potentials for fold assessment. *Protein Sci*. 2002;430:430–448.
50. Park B, Levitt M. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J Mol Biol* 1996;258:367–392.
51. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 1997;268:209–225.
52. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol* 2000;300:171–185.
53. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* 1998;275:895–916.
54. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* 2001;44:223–232.
55. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins* 2000;41:40–46.
56. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. In: Altman RB, Dunker AK, Hunter, Lauderdale K, Klein TE, editors. *Pacific Symp. Biocomputing*. New York: World Scientific; 2000. p 119–130.
57. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
58. Fischer D, Elofsson A, Rice D. W, Eisenberg D. Assessing the performance of inverted protein folding methods by means of an extensive benchmark. In: *Proceeding of the First Pacific Symposium on Biocomputing*. 1996;300–318.
59. Siew N, Elofsson A, Rychlewski L, Fischer D. Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* 2000;16:776–785.
60. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Šali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 2000;29:291–325.
61. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 1999;287:797–815.
62. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol* 2000;297:1003–1013.
63. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. *J Mol Biol* 2000;295:613–625.
64. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 2001;310:243–257.
65. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
66. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Livebench-1: Large-scale automated evaluation of protein structure prediction servers. *Protein Sci* 2001;10:352–361.
67. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19:1589–1591.
68. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. *J Bioinformatics Computational Biol* 2003;1:95–117.
69. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins* 2003;51:434–441.