PROTEINS: Structure, Function, and Bioinformatics 58:321–328 (2005)

Fold Recognition by Combining Sequence Profiles Derived From Evolution and From Depth-Dependent Structural Alignment of Fragments

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, Buffalo, New York

ABSTRACT **Recognizing structural similarity without significant sequence identity has proved to be a challenging task. Sequence-based and structure-based methods as well as their combinations have been developed. Here, we propose a fold-recognition method that incorporates structural information without the need of sequence-to-structure threading. This is accomplished by generating sequence profiles from protein structural fragments. The structure-derived sequence profiles allow a simple integration with evolution-derived sequence profiles and secondary-structural information for an optimized alignment by efficient dynamic programming. The resulting method (called SP^3) is found to make a statistically significant improvement in both sensitivity of fold recognition and accuracy of alignment over the method based on evolution-derived sequence profiles alone (SP) and the method based on evolution-derived sequence profile and secondary structure profile (SP^2). SP^3 was tested in SALIGN benchmark for alignment accuracy and Lindahl, PROSPECTOR 3.0, and LiveBench 8.0 benchmarks for remote-homology detection and model accuracy. SP^3 is found to be the most sensitive and accurate single-method server in all benchmarks tested where other methods are available for comparison (although its results are statistically indistinguishable from the next best in some cases and the comparison is subjected to the limitation of time-dependent sequence and/or structural library used by different methods.). In LiveBench 8.0, its accuracy rivals some of the consensus methods such as ShotGun-INBGU, Pmodeller3, Pcons4, and ROBETTA. SP^3 fold-recognition server is available on http://theory.med.buffalo.edu. Proteins 2005;58:321–328.**

Key words: fold recognition; protein threading; protein structure prediction; sequence profile

INTRODUCTION

Fold recognition refers to recognition of structural similarity without significant sequence identity. One way to detect structural similarity is to identify remote sequence homology via sequence comparison. Advances have been made from the pairwise[1–7] to multiple[8–12] sequence comparison, from sequence-to-sequence, sequence-to-profile[8,9,13] to profile-to-profile comparison.[12,14–17] Here, a sequence profile is a position-dependent probability of amino acid residues usually obtained from multiple sequence alignment.[14] Several recent works compared the different techniques for profile-profile alignments.[17–19]

Another way to detect structural similarity is to take full advantage of known protein structures. For example, the sequence-to-structure threading[1] assesses the compatibility of a sequence with each known structure by a pairwise score function or single-body structural profile.[20–25] (For recent reviews, see Refs.[26–30].) More recent work attempt to optimally combine the sequence and structure information for a more accurate/sensitive fold recognition.[7,31–43] Most focused on combining sequence information with threading techniques.

One intuitive approach to incorporate structural information is structural alignment.[14] Application of structural alignment to fold recognition has been mostly limited to the derivation of substitution matrices.[44–47] The direct incorporation of sequence profiles generated from structural alignment, however, does not appear to be useful for remote homology detection.[48–51] For example, Gough et al.[48] found that hidden Markov models (HMM) generated from structural alignment yielded poorer results than HMMs generated independently. Tang et al.[42] showed that the combination of sequence profiles derived from structural alignments for protein-core regions with the sequence profiles from sequence alignment and secondary structural profiles does not further improve fold-recognition sensitivity by profile-profile alignment. This highlights the difficulty of harnessing structural information in a combined approach for optimal fold-recognition alignment.[38,52] In fact, recently completed LiveBench[53] 8 indicates that all top four performers of the fold-recognition servers of single methods are sequence-based profile-profile alignment methods (BasD/mBas/BasP, SFST/

Grant sponsor: the National Institutes of Health; Grant numbers: R01 GM 966049 and R01 GM 068530.

*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 18 May 2004; Accepted 22 July 2004

Published online 1 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20308

STMP, FFAS03,[15] and ORFeus/ORFeus2[54]). So far, the most successful combinational uses of structural and sequence information are those consensus methods[55] that make predictions based the results of single methods. One example is ShotGun-INBGU which uses the ShotGun method[56] to create alternative consensus models from the INBGU components.[57]

A possible source of the problem associated with direct use of protein–protein structural alignment for fold recognition is that alignment may not have a unique solution even for the core regions.[58] Moreover, loop regions often do not have a meaningful structural alignment. Thus, we propose to use fragments rather than whole proteins for structural alignment. The use of fragments, however, loses the information on the environment surrounding the fragments. This is remedied by incorporating the information on the depth of residues[59] from protein surface in measuring fragment similarity. The sequence profile (SP) derived from depth-dependent structure alignment of fragments is used along with evolution-derived sequence profile (SP) and the secondary structure profile (SP). The new fold-recognition method is called SP^3. It is shown that the new structure-derived sequence profile improves not only the alignment accuracy but also the recognition sensitivity over the fold-recognition methods SP (sequence profiles only) and SP^2 (sequence plus secondary structure profiles). The performance of SP^3 on fold recognition is outstanding in several testing benchmarks.

METHOD

The fold-recognition method proposed here performs a profile–profile comparison not only between the evolutionary-derived sequence profile of a query sequence and that of a template sequence but also between the query sequence profile and template sequence profile derived from the structures of fragments of the template.

Sequence Profile From Depth-Dependent Structural Alignment of Fragments

Structure-based sequence profile for a given template structure is generated as follows. Each template structure is divided into structural fragments with a sliding window size of I_w amino acid residues (i.e., from 1 to I_w, 2 to $I_w + 1$, 3 to $I_w + 2$, . . .). Each structural fragment is compared with same-size fragments contained in a large database of protein structures. We measure fragment similarity by both fragment–structural and protein–environment similarities. This is accomplished by using a similarity score that weights equally the root-mean-squared distance between the two fragment structures and an exponential function that characterizes the difference between their solvent exposures. The similarity score between a template fragment and a same-size fragment from structural database is given by the equation

$$S_{str} = d_{rmsd}^2 + w_d \sum_{j \in window} [\exp(-D_j^{template}/2.8) - \exp(-D_j^{database}/2.8)]^2 \quad (1)$$

where d_{rmsd} is the root-mean-squared distance between the two fragment structures, w_d is a weight factor, $D_j^{template}$ and $D_j^{database}$ are the depth of the fragment residue j from the surface of the template structure and the depth of the corresponding residue from the surface of the protein structure in the structural database, respectively, and the summation ($j = 1,2,\ldots,I_w$) is over all the positions in the fragments. The depth of residue from surface is calculated by the method described in detail in Chakravarty and Varadarajan.[59] The basic idea is to calculate the average shortest distance of a residue from solvent water molecules. Here, we set $w_d = 10$ to make the two terms comparable in magnitude. The depth (in Å) is scaled by 2.8 Å, the approximate size of a water molecule. We use a window size of 9 ($I_w = 9$), the same size used by Simons et al. to build a structural fragment library.[60] The structure database of fragments is made of 1011 nonhomologous (less than 30% homology) proteins with resolution < 2 Å that was collected by the program PISCES (http://chaos.fccc.edu/research /labs/dunbrack/culledpdb.html) .[61]

After a template fragment is compared against all same-size fragments in the structural database, the top 25 fragments ranked by the similarity score S_{str} are retained for profile construction. (One can use more or less fragments for this purpose. We did not attempt to optimize this parameter.) Because each residue position (except near the C or N terminus) associates with 9 template fragments and each template fragment has top 25 similar fragments from the structural database, there are 225 (9×25) sequences that can be used to compute the frequencies of the 20 amino acid residues at a given residue position (less for the four residues near the C or N terminus). The frequency profiles obtained are normalized so that the summation over 20 types of amino acid residues is equal to 1. Here and hereafter, this structure-derived frequency profile at sequence position j will be labeled as $F_{template}^{struc}(j)$.

There exists a possible bias to the sequence profiles generated by the above method. This happens when there are several hits from the same residue in a single sequence to the same single profile position. The maximum number of such multiple hits is nine. This number is small, relative to the total of 225 sequences used in profile generation. The possible effect of such bias on fold recognition is not tested and will be a subject of further study.

Sequence Profile Derived From Sequence Library

All sequence-derived profiles are constructed by using PSIBLAST. This is done with three iterations of searching against nonredundant (NR) sequence database (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz). As in PSIPRED,[62] the database was filtered to remove low-complexity regions, transmembrane regions, and coiled-coil segments.

Secondary Structures

The secondary structure of query sequence is predicted by a built-in simplified PSIPRED method.[43] The secondary structures of templates were obtained by H-bonds (DSSP-like) criteria.[63] Three states (helix, strand, coil) were used for all secondary structures.

The Alignment Score and the Alignment Algorithm

The alignment score for aligning query sequence position i with the template sequence position j is given by the equation

$$S(i,j) = -(1 - w_{struc})F^{seq}_{query}(i) \cdot M^{seq}_{template}(j) - w_{struc}F^{struc}_{template}(j) \cdot M^{seq}_{query}(i) - w_{2ndary}\delta_{si,sj} + s_{shift} \quad (2)$$

where $F^{seq}_{query}(i)$ is the sequence-derived frequency profile of the query sequence, $M^{seq}_{template}(j)$ is the sequence-based log odd profile (position-specific substitution matrix as in PSIPRED) of the template, $F^{struc}_{template}(j)$ is the structure-derived frequency profile of the template, $M^{seq}_{query}(i)$ is the sequence-derived, log odd profile (position-specific substitution matrix) of the query sequence, s_{shift} is a to-be-determined constant shift, w_{struc} and w_{2ndary} are two weight parameters for structure-derived sequence profiles and secondary structure profiles, respectively, and $\delta_{si,sj}$ is a simple function of the secondary structure element si of the query at sequence position i and sj of the template at sequence position j.

$$\delta_{si,sj} = \begin{cases} 1, & si = sj, \\ -1, & si \neq sj \end{cases}. \quad (3)$$

Because we use the Smith-Waterman local alignment algorithm to align profiles (see below), s_{shift} is used to avoid the alignment of unrelated regions. A position-dependent gap penalty is employed. No gaps are allowed if $si = sj = \alpha$ or $si = sj = \beta$. The gap opening (w_0) and gap extension (w_1) penalties are applied to other regions. Finally, a local–local dynamic programming method[64] is used to optimize the score that matches the query profiles with template profiles. Note that the optimization of alignment is to minimize the total alignment score due to the negative signs in equation 2.

Ranking Templates and Model Assessments

The following empirical method is used for ranking. First, a difference raw score is calculated. A difference raw score ΔS is the difference between the alignment raw score (S) and the reverse alignment raw score S_r in which the alignment is made with the reversed query sequence.[9] If there is structural similarity between first two models ranked by ΔS (defined as nonzero MaxSub score[65]), all models will be ranked by ΔS. Otherwise, two normalized scores (S_{n1} and S_{n2}) and their corresponding Z-scores (Z_1 and Z_2) are calculated. All templates will be reranked by either Z_1 or Z_2 depending on which Z-score for the first model ranked by S_{n1} and S_{n2} is greater. Here, S_{n1} is the raw score normalized by the full alignment length between the query and template sequences (including all gaps). The full alignment length is close to the length of longer sequence of the query and the template sequences. S_{n2}, on the other hand, is the raw score normalized by the alignment length excluding query end gaps. $Z_1(i) = (S_{n1}(i) - S^{ave}_{n1})/S^{sd}_{n1}$ and $Z_2(i) = (S_{n2}(i) - S^{ave}_{n2})/S^{sd}_{n2}$, where superscripts *ave* and *sd* denote the average and standard deviation of normalized score for all the templates.

The results of fold-recognition alignment are used to build C_α models based on template structure. The models are then assessed by the MaxSub score between the model and the known native structure. MaxSub score[65] between the predicted (model) structure and the native structure is a measure of similarity between 0.0 (no similarity) and 1.0 (perfect similarity). The value is calculated by searching the largest subset of well-superimposed residues (≤ 3.5 Å). We use MaxSub score because it is the official evaluation method used in the CAFASP (Critical assessment of fully automated protein structure prediction methods) experiments (http://www.cs.bgu.ac.il/dfischer/cafasp1/cafasp1.html).

A Note on Comparison

It should be emphasized that comparison made between this work and other published work is not a strict one. This is because it is impossible to have an exact comparison between the methods that use the time-dependent sequence and/or structural libraries.

RESULT

Optimization of Parameters Using ProSup Benchmark

For the SP^3 fold-recognition method, there are five adjustable parameters (w_0, w_1, w_{2ndary}, w_{struc}, and s_{shift}). For comparison, we also optimized the parameters for the SP (sequence profiles only, $w_{2ndary} = w_{struc} = 0$) and the SP^2 method (sequence profiles and secondary-structure profiles, $w_{struc} = 0$). For SP, position-dependent gap penalty is not used because secondary-structure information is not available. We use the ProSup benchmark set for optimizing all the parameters. This benchmark was prepared by Sippl's group[66] 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. The optimization is started with random values for all the parameters. Then, all parameters are optimized by grid search sequentially. The optimization procedure is stopped when iterations do not improve the alignment accuracy. This procedure may lead to local minimum. Thus, several independent optimizations based on different initial values were made. The final optimized parameter sets are (6.6, 0.58, −0.9) for (w_0,w_1, s_{shift}) in SP, (7.8, 0.18, 0.73, −1.30) for (w_0, w_1, w_{2ndary}, s_{shift}) in SP^2, and (5.2, 0.38, 0.38, 0.50, −1.15) for (w_0, w_1, w_{2ndary}, w_{struc} and s_{shift}) in SP^3.

Table I compares the performance of SP, SP^2, and SP^3 methods. It shows that the secondary structure information leads about 7% leap in alignment accuracy whereas the structure-derived profile introduces an additional 2–3% improvement. The performance of several other methods is also listed along with SP methods. Their results are used as a reference because ProSup was a testing benchmark for those methods. Nevertheless, a significantly better performance than the next best (SPARKS or PROSPECT

TABLE I. The Average Alignment Accuracy for ProSup Benchmark per Pair of Proteins[†]

Method	Accuracy[a]	±4 residues[b] (%)
SP[c]	55.9	72.4
SP[2c]	62.9	79.1
SP[3c]	65.3	82.2
FASTA[d]	31.4	
Sequence[d]	34.1	
PSI-BLAST[e]	35.6	
Threading[f]	48.0	
PROSPECT II[g]	57.7	75.8
SPARKS[e]	57.2	78.5

[†]127 pairs of proteins aligned by the ProSup program.
[a]Accuracy defined by one-to-one match given by the method and the benchmark.
[b]The accuracy defined by the match within four residues from the one-to-one match.
[c]This work.
[d]Pairwise sequence comparisons. Results from Sippl et al.[66]
[e]From Zhou and Zhou[43] (3 iteration and 0.001 E-value as for generating sequence profiles).
[f]From Sippl et al.[66] This result may not reflect the accuracy of the current version of the threading method.
[g]From Kim et al.[41]

TABLE II. The Average Alignment Accuracy Assessed by CE Overlap in SALIGN Benchmark per Pair of Proteins (200 Protein Pairs)

Method	CE overlap[a] (%)
BLAST[b]	26.1
SEA[b]	49.2
SAM[b]	48.4
LOBSTER[b]	49.9
SALIGN[b]	56.4
SPARKS[c]	53.1
SP[3d]	56.6
SP[d]	52.0
SP[2d]	54.5

[a]The percentage of aligned positions that were identical to those in the structure-based CE alignment.
[b]From Marti-Renom et al.[17] SEA (Segment alignment),[74] SAM package,[75] LOBSTER.[76]
[c]Using the method developed in Zhou and Zhou.[43]
[d]This work by using the first chain as query and second chain as template.

II) (a 7% improvement in terms of one-to-one match) indicates that the SP^3 method is promising to provide a more accurate fold-recognition alignment.

Test Set 1: SALIGN Benchmark

To test the alignment accuracy, we used the SALIGN benchmark.[17] This benchmark contains 200 selected pairs of proteins that were structurally aligned by CE program.[67] The selected structure pairs have an average pair sharing 20% sequence identity and 65% of structurally equivalent C_α atoms superposed with an rmsd of 3.5 Å.[17] This benchmark can be considered as an independent benchmark from the ProSup benchmark because sequence identities between any pairs of proteins from the two benchmarks are less than 40% at least for one of the proteins in the pair. Alignment accuracy for the SALIGN benchmark is assessed by calculating the fraction of the alignment that is the same as the alignment obtained from the CE program (CE overlap).

The results on CE overlap by SP methods are compared with those obtained by several other methods tested by Marti-Renom et al.[17] in Table II. The SP3 method gives slightly better success rate than the best profile–profile alignment method SALIGN by Marti-Renom et al.[17] who compared 13 different alignment protocols and several established methods. This is remarkable considering the fact that SP^3 was trained by a different structural alignment program called ProSup whereas SALIGN was trained by and tested on the CE alignment. One-to-one matches defined by ProSup and that by the CE alignment are very different. For example, if the alignment accuracy for one-to-one match in ProSup benchmark is defined by CE alignment, the accuracy of alignment in ProSup benchmark given by SP^3 will be reduced significantly from 65.3% to 49.5%.

The change of alignment accuracy from SP, SP^2, to SP^3 in SALIGN benchmark is not as dramatic as the change in the training ProSup benchmark (Table I). The former is only 4.6% compared to 9.4% in the latter. The 4.6% improvement is more-or-less equally distributed between SP and SP^2 and between SP^2 to SP^3. Nevertheless, the improvement of alignment accuracy from SP to SP^3 is reproduced despite different structural alignment programs used in SALIGN and ProSup benchmarks. One important question is whether or not the differences between the performances of SP methods are statistically significant. Simple student t-test[68] based on the average and standard deviation of the differences indicates that both differences (2.5% and 2.1%) are significant at 95% confidence level.

Test Set 2: Lindahl Benchmark for Fold-Recognition Sensitivity

The Lindahl set[52] 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 SP^3 are compared with several well-established methods in Table III. We stress that the comparison only serves as an approximate guide because the sequence database available for previous methods are smaller than the one used in SP^3. It shows that SP^3 is the most sensitive method to detect structural similarity on the fold and superfamily levels for the first rank among the eleven methods listed. For example, compared to the popular PSI-BLAST, SP^3 is 10%, 28%, and 25% more sensitive in recognizing the member of same family, superfamily, and fold, respectively.

Table III shows that the performances of SP^3, SPARKS, and PROSPECT II are very similar to each other. One is

TABLE III. Performance of Fold Recognition for Lindahl Benchmark[52]

	Family only		Superfamily only		Fold only	
Method	Top 1 (%)	Top 5 (%)	Top 1 (%)	Top 5 (%)	Top 1 (%)	Top 5 (%)
PSI-BLAST[a]	71.2[b]	72.3	27.4	27.9	4.0	4.7
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
THREADER[a]	49.2	58.9	10.8	24.7	14.6	37.7
FUGUE[a]	82.2	85.8	41.9	53.2	12.5	26.8
RAPTOR[c]	75.2	77.8	39.3	50.0	25.4	45.1
PROSPECT II[d]	84.1	88.2	52.6	64.8	27.7	50.3
SPARKS[e]	81.6	88.1	52.5	69.1	24.3	47.7
SP^3[f]	81.6	86.8	55.3	67.7	28.7	47.4
SP[f]	81.3	87.2	51.8	65.0	20.2	35.8
SP^2[f]	82.5	87.0	52.5	67.1	24.9	40.2

[a]From Shi et al.,[47] the upgraded version of these methods may perform better than the above results.
[b]The percentage in each cell is the fraction of correctly recognized match of proteins in the same fold, super family, family as first rank or within top five rank of the template.
[c]From Xu et al.[77] Also see comment in footnote a.
[d]From Kim et al.[41] Also see comment in footnote a.
[e]From Zhou and Zhou.[43]
[f]This work.

only a few percentage (1–3%) better than the other. This sensitivity test, however, may not reflect the true sensitivity or accuracy because the result is based on somewhat subjective SCOP classification.[69] To address this question more quantitatively, we calculated the MaxSub score between the model built from first-ranked template and the known native structure. It is found that the total number of recognized proteins (MaxSub > 0.01, a measure of sensitivity[53]) increases from 611 in SPARKS to 665 in SP^3 (a 8.6% improvement) whereas the total MaxSub score (a measure of model accuracy) increases from 325.85 in SPARKS to 349.20 in SP^3 (a 7.2% increment). (Results reported here are based on C_α models generated from alignment.)

Within SP methods, SP^3 improves over SP^2 and SP at both fold and superfamily levels. For example, 3–4% improvements of SP^3 over SP^2 are observed for the first ranking models at fold and superfamily levels. At the family level, SP, SP^2, and SP^3 have similar sensitivity (between 81.3%–82.5%). Improvement in recognition sensitivity is further confirmed based on MaxSub score. The number of the first-ranked models with MaxSub > 0.01 is 606 for SP, 638 for SP^2, and 665 for SP^3 while the total MaxSub score increases from 328.6, 340.8, to 349.2.

Test Set 3: PROSPECTOR 3.0 Benchmark

The PROSPECTOR 3.0 Benchmark is a large benchmark of 1479 targets and 3825 templates. Each target also has a list of templates excluded due to their sequence similarity to the target. (Two targets 1ddqE and 1f5yA are removed from the original 1481 targets due to lack of a template exclusion list for these two targets.) This benchmark is an updated version from what was published.[70]

Table IV compares the results of SPARKS, SP^3 with that of PROSPECTOR 3.0. Based on MaxSub score, SP^3 is about 15% and 9% more sensitive than PROSPECTOR 3.0 and SPARKS, respectively, and about 16% and 14% more accurate in model accuracy than PROSPECTOR 3.0 and SPARKS, respectively. SP^3 also shows significant improvement (3% and 6%, respectively) in both sensitivity and accuracy over SP^2 and SP. (Comparison of PROSPECTOR 3.0 and SPARKS in Skolnick et al.[70] suggested that SPARKS is not as accurate as PROSPECTOR 3.0. The conclusion is different from what we obtained here. This is in part due to different definition of model accuracy and the use of different sequence library in their implementation of SPARKS.)

TABLE IV. The Performance of Various Methods on the PROSPECTOR 3.0 Benchmark (1479 Targets and 3825 Templates) Based on First-Ranked Models

Method	Sensitivity[a]	Total MaxSub score
PROSPECTOR3.0[b]	925	520.1
SPARKS[c]	979	529.0
SP^3[d]	1066	601.9
SP[d]	1000	564.6
SP^2[d]	1034	583.2

[a]The numbers of targets with MaxSub score > 0.01.
[b]Calculated based on the models downloaded from the webpage: http://www.bioinformatics.buffalo.edu/threadingbenchmark_2
[c]Method of Zhou and Zhou.[43]
[d]This work.

Test Set 4: LiveBench 8

The template library for SP^3 was built as the library for SPARKS.[43] This was done by using the 40% representative domains of SCOP 1.61. The entire chains of multiple-domain proteins are contained in the library. The library was then updated with new proteins released after SCOP

TABLE V. Performance for the 172 LiveBench 8 Targets[†]

Method	Sensitivity[a]	Total MaxSub score	Specificity[b]
SPARKS[c]	99	38.33	79.5
SAM-T02[c]	101	39.16	95.4
ORFeus[c]	105	39.79	92.6
FFAS03[c]	105	40.01	83.8
STMP[c]	107	40.47	88.5
BasD[c]	112	41.91	100.2
ShotGun-INBGU[c]	107	43.30	92.1
SP^{3d}	120	42.24	94.0
SP^{d}	106	37.18	87.8
SP^{2d}	116	39.51	94.2

[†]Only Individual Servers that have higher total MaxSub scores than SPARKS are listed below.
[a]Sensitivity is the number of targets whose first-ranking models with a MaxSub score of greater than 0.01.
[b]The specificity is define as the average number of recognized proteins that have scores better than 1–10 false positives.
[c]Results from LiveBench server (http://BioInfo.PL).
[d]This work.

1.61 if they have less than 40% sequence identity with the sequences already in the library. (This was done by protein sequence culling server PISCES[61]).

Unlike SPARKS, SP^3 was not directly involved in LiveBench test. Instead, SP^3 used the exactly same template library of SPARKS for each target to "simulate" live testing. Here, all SP^3 results are based on C_α models of aligned residues. Only results of top single servers are shown in Table V. Three parameters are used to compare different methods. The accuracy is characterized by the total MaxSub scores. Sensitivity is defined as the number of targets whose first-ranking models have a MaxSub score of greater than 0.01. Specificity is defined as the average number of recognized proteins that have scores better than 1–10 false positives. For 172 Livebench 8 targets, SP^3 has the highest sensitivity and second highest in total MaxSub score (behind ShotGun-INBGU). Among the top single servers, ShotGun-INBGU is the only consensus method that provides consensus model from the INBGU components.[57] Others are single-method servers. It is remarkable that the accuracy of SP^3 is comparable to or better than some of the consensus methods such as ShotGun-INBGU, Pmodeller3 (Total MaxSub=42.77),[71] Pcons4 (42.53),[71] and ROBETTA (41.79)[72] although the model accuracy of SP^3 is still about 10% less than the best consensus methods in LiveBench (e.g., ShotGun on 5 and ShotGun on 3). However, the sensitivity of SP^3 (120 in 172 proteins) is the highest for all the methods (more than 40 servers) tested in LiveBench 8. The next best is 119 by a consensus method called 3D-JuryB-single. It should be cautioned, however, that only the difference between SPARKS and SP^3 is statistically significant based on student-t test. The differences of performance in LiveBench 8 between all other methods (SAM-T02, ORFeus, FFAS03, STMP, BasD, ShotGun-INBGU and SP^3) are statistically insignificant.

DISCUSSION

This paper describes a new method to combine sequence and structural information for an optimized fold-recognition alignment. The method employs the sequence profile generated from structural and residue-depth alignment of fragments rather than the commonly used structural alignment of whole proteins. The combination of this structure-based sequence profile with evolution-based sequence profile and secondary structural profile yields the method called SP^3. Unlike previous attempts to use sequence profile generated from structural alignment,[48,42] the new structure-derived sequence profile leads to a statistically significant improvement in both accuracy and sensitivity of fold recognition as demonstrated by testing on several large benchmarks. [Structural alignment, however, was used successfully to generate score functions (but not sequence profiles) for fold recognition, as in, for example, 3D-PSSM[37]]. Improvement of SP^3 in model accuracy (based on total MaxSub score) over SP^2 is 2% in Lindahl benchmark, 3% in PROSPECTOR 3.0 benchmark, and 7% in LiveBench 8.0. Improvement of SP^3 in recognition sensitivity (MaxSub $>$ 0.01) over SP^2 is 4% in Lindahl benchmark, 3% in PROSPECTOR 3.0 benchmark, and 3% in LiveBench 8.0. A simple student t-test of above results indicates that both improvements in accuracy and sensitivity are statistically significant.

The sequence profile generated from structural fragments is shown to be successful in improving detection of remote homology. The use of fragments has following advantages over the use of whole protein for structural alignment. First, there is a sufficient coverage for all possible structures of short fragments in the existing structures in protein data bank. Du et al.[73] showed that there is 96% success rate for finding two similar nine-residue fragments within 1 Å RMSD. The large number of fragments contained in protein data bank leads to a statistically significant sequence profile. In contrast, sequence profiles generated from structural alignment of whole proteins[48,42] require that all proteins have a sufficient number of structurally similar proteins with low sequence identity—a condition that is difficult to meet. Second, the use of fragments allows the generatation of a reliable sequence profile for all regions of a protein. On the other hand, many regions (loop regions, in particular) are not aligned in structural alignment of whole proteins. Third, unlike structural alignment of proteins,[58] structural alignment of fragments is more likely to have a unique solution because their structural topologies are relatively simple.

One unique feature of the SP^3 method is that alignment of two fragments is not only characterized by their structural difference (RMSD) but also by their positions from solvent (residue depth). To assess if depth plays a significant role in accuracy of fold recognition, we performed the parameter optimization by switching off the depth score in equation 1 (i.e., $w_d = 0$). The resulting SP^3 leads to an alignment accuracy of 63.7% for one-to-one match in ProSup benchmark. This accuracy is somewhat in between 62.9% (SP^2) and 65.3% (SP^3). Thus, the use of depth in

structural alignment contributes to the improved accuracy of alignment in SP^3.

The results of LiveBench 8 indicate no statistically significant difference among the top single-server methods. This could signal a common bottleneck reached by various fold-recognition methods. It is more likely, however, caused by the small number of targets (172) in LiveBench 8. This is because LiveBench 7 with 115 targets yielded a very different ranking. For example, BasD and ORFeus were ranked behind SPARKS in LiveBench 7. Only three methods (ShotGun-INBGU, STMP, FFAS3) are both in top six single servers for LiveBench 7 and LiveBench 8. This suggests the importance of using large benchmarks for a statistically significant comparison between different methods.[18,19,70,68]

ACKNOWLEDGMENTS

We gratefully thank Professors Jeff Skolnick and Yang Zhang for providing us the preprint and benchmark of PROSPECTOR 3.0 prior to its publication, Professor Roland Dunbrack for his preprint on profile scoring. This work was supported by NIH (R01 GM 966049 and R01 GM 068530), a grant from HHMI to SUNY Buffalo and by the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Y. Z. was also supported in part by a two-base fund from the National Science Foundation of China.

REFERENCES

1. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
2. Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. Meth Enzymol 1983;91:524–545.
3. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 1988;85:2444–2448.
4. Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment tool. J Mol Biol 1990;215:403–410.
5. Vingron M, Waterman MS. Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol 1994;235:1–12.
6. Qian B, Goldstein RA. Optimization of a new score function for the generation of accurate alignments. Proteins 2002;48:605–610.
7. Teodorescu O, Galor T, Pillardy J, Elber R. Enriching the sequence substitution matrix by structural information. Proteins 2004;54:41–48.
8. Altschul SF, Madden TL, Schäffer 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.
9. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
10. Henikoff S, Henikoff JG. Amino acid substitutes matrices from protein blocks. Proc Natl Acad Sci 1992;89:10915–10919.
11. Bailey TL, Gribskov M. Score distributions for simultaneous matching to multiple motifs. J Comput Biol 1997;4:45–59.
12. Koretke KK, Russell RB, Lupas AN. Fold recognition from sequence comparisons. Proteins 2001;Suppl 5:68–75.
13. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
14. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
15. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
16. 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.
17. Marti-Renom MA, Madhusudhan M, Sali A. Alignment of protein sequences by their profiles. Protein Sci 2004;13:1071–1087.
18. Mittelman D, Sadreyev R, Grishin N. Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. Bioinformatics 2003;19:1531–1539.
19. Wang G, Dunbrack RL Jr. Scoring profile-profile sequence alignments. Protein Sci 2004;13:1612-1626.
20. 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.
21. 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.
22. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
23. Abagyan R, Frishman D, Argos P. Recognition of distantly related proteins through energy calculations. Proteins 1994;19:132–140.
24. Murzin AG, Bateman A. Distance homology recognition using structural classification of proteins. Proteins 1997;Suppl. 1:105–112.
25. Xu Y, Xu D. Protein threading using PROSPECT: Design and evaluation. Proteins 2000;40:343–354.
26. Jones DT. Progress in protein structure prediction. Curr Opin Struct Biol 1997;7:377– 387.
27. Torda AE. Perspectives in protein-fold recognition. Curr Opin Struct Biol 1997;7:200–205.
28. David R, Korenberg MJ, Hunter IW. 3D-1D threading methods for protein fold recognition. Pharmacogenomics 2000;1:445–455.
29. Sippl MJ, Lackner P, Domingues FS, Prlić A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001; Suppl 5:55–67.
30. 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.
31. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42:319–331.
32. Yi TM, Lander ES. Recognition of related proteins by iterative template refinement (ITR). Protein Sci 1994;3:1315–1328.
33. Elofsson A, Fischer D, Rice DW, Le Grand SM, Eisenberg D. A study of combined structure/sequence profiles. Fold Des 1996;1: 451–461.
34. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
35. Rost B, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997; 270:471–480.
36. 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.
37. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
38. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
39. 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.
40. 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.
41. Kim D, Xu D, Guo J, Ellrott K, Xu Y. PROSPECT II: Protein structure prediction program for the genome-scale. Protein Eng 2003;16:641–650.
42. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B. On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 2003;334:1043–1062.
43. Zhou H, Zhou Y. Single-body knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55:1005–1013.
44. Ogata K, Ohya M, Umeyama H. Amino acid similarity matrix for homology derived from structural alignment and optimized by the Monte Carlo method. J Mol Graph Model 1998;16:178–189.
45. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution

matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–550.
46. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.
47. 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.
48. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. J Mol Biol 2001;313.
49. Panchenko AR, Bryant SH. A comparison of position-specific score matrices based on sequence and structure alignments. Protein Sci 2002;11:361–370.
50. Gough J, Chothia C. Superfamily:HMMs representing all proteins of known structure. Scop sequence searches, alignments, and genome assignments. Nucleic Acids Res 2002;30:268–272.
51. Griffiths-Jones S, Bateman A. The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. Bioinformatics 2002;18:1243– 1249.
52. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. J Mol Biol 2000;295:613–625.
53. 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.
54. Ginalski K, Pas J, Wyrwicz LS, vonGrotthuss M, Bujnicki JM, Rychlewski L. ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 2003;31:3804–3807.
55. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)—Round V. Proteins 2003;53:334–339.
56. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
57. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. In: Altman RB, Dunker AK, Hunter L, Klein TE, editors. Pacific Symp. Biocomputing. New York: World Scientific; 2000. p 119–130.
58. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338.
59. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Strucure Fold Des 1999;15:723–732.
60. 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.
61. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
62. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
63. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
64. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
65. 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.
66. 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.
67. Shindyalov IN, Bourne P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
68. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. Structure 2002;10:435–440.
69. 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.
70. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004;55:502-518.
71. Wallner B, Fang H, Elofsson A. Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. Proteins 2003;53: 534–541.
72. Chivian D, Kim DE, Malmström L, Bradley P, Robertson T, Murphy P, Strauss C. E, Bonneau R, Rohl CA, Baker D. Automated prediction of casp-5 structures using the robetta server. Proteins 2003;53:524–533.
73. Du P, Andrec M, Levy RM. Have we seen all structures corresponding to short protein fragments in the protein data bank? An update. Protein Eng 2003;16:407–414.
74. Ye Y, Jaroszewski L, LiW, Godzik A. A segment alignment approach to protein comparison. Bioinformatics 2003;19:742.
75. Hughey R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput Appl Biosci 1996;12:95–107.
76. Edgar RC, Sjolander K. Simultaneous sequence alignment and tree construction using hidden markov models. Pac Symp Biocomput 2003;180–191.
77. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol 2003;1:95–117.