Fold recognition by concurrent use of solvent accessibility and residue depth

Song Liu,[1] Chi Zhang,[1,2] Shide Liang,[1,2] and Yaoqi Zhou[1,2]*

[1] Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology and Biophysics, State University of New York at Buffalo, Buffalo, New York 14214

[2] Indiana University School of Informatics, Indiana University-Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202

ABSTRACT

Recognizing the structural similarity without significant sequence identity (called fold recognition) is the key for bridging the gap between the number of known protein sequences and the number of structures solved. Previously, we developed a fold-recognition method called SP³ which combines sequence-derived sequence profiles, secondary-structure profiles and residue-depth dependent, structure-derived sequence profiles. The use of residue-depth-dependent profiles makes SP³ one of the best automatic predictors in CASP 6. Because residue depth (RD) and solvent accessible surface area (solvent accessibility) are complementary in describing the exposure of a residue to solvent, we test whether or not incorporation of solvent-accessibility profiles into SP³ could further increase the accuracy of fold recognition. The resulting method, called SP⁴, was tested in SALIGN benchmark for alignment accuracy and Lindahl, LiveBench 8 and CASP7 blind prediction for fold recognition sensitivity and model-structure accuracy. For remote homologs, SP⁴ is found to consistently improve over SP³ in the accuracy of sequence alignment and predicted structural models as well as in the sensitivity of fold recognition. Our result suggests that RD and solvent accessibility can be used concurrently for improving the accuracy and sensitivity of fold recognition. The SP⁴ server and its local usage package are available on http://sparks.informatics.iupui.edu/SP4.

INTRODUCTION

An effective method for protein structure prediction is homology or comparative modeling, which predicts the structure of a query sequence based on the similarity of the sequence to the sequences with known three-dimensional structures. Fold recognition, on the other hand, goes beyond sequence similarity by attempting to recognize structurally related sequences in the absence of significant sequence identity. An accurate method for fold recognition is essential for achieving the goal of structural genomics project[1,2] because proteins adopt a limited number of unique structural folds many of which may have already been solved.[3,4]

One approach to improve the accuracy of fold-recognition is to improve the accuracy of a single method for aligning remote homologs. Advances have been made from profile-based alignment techniques,[5–15] to a combination of sequence profile and structural information (from either threading or structure-profile scores).[16–26] Methods of machine-learning-based information retrieval have also been developed.[18,27] For recent review or evaluation, see Refs. 28–34.

Another approach is to integrate several existing methods for an improved ability of detecting structural homologs. This type of consensus methods (or meta-servers), collects models from other single servers and uses them to generate consensus predictions.[35–39] The emergence of consensus methods has led to a significant improvement in the quality of predicted structures, as monitored by recent CASP, CAFASP, and LiveBench competitions.[40] However, these methods rely on the availability and accuracy of individual methods. Thus, developing or improving individual methods, albeit challenging, is crucial[41] to further advance the accuracy of fold recognition.

In searching for an improved fold-recognition method, we focus on the effect of solvation since the driving force of protein folding is water mediated interaction among amino acid residues. The most widely-used geomet-

rical parameter to quantify the effect of solvation is solvent accessibilities (SA) of amino acid residues. SA of an amino acid residue in a protein is defined here as the ratio of the solvent accessible surface area of the residue in a folded protein structure to that in the unfolded structure.[42] SA of a residue equals to 0% (or 100%) if the residue is fully buried from (or fully accessible by) solvent. In general, hydrophobic residues are less solvent accessible than hydrophilic residues.

The SAs of residues at different sequence positions can be calculated analytically for any template structures. This information together with the propensity score of residual burial tendency (mainly derived from either representative protein structures or polar-to-nonpolar solvent transfer energy), has been widely used in the sequence-to-structure matching scores in many methods.[17,18,43,44] Predicted SAs of query sequences,[45,46] like predicted secondary structures,[47–49] are also employed,[50–53] although SA is less conserved than secondary structures and thus more difficult to predict.[45,46] It is expected that SA information might be useful in sequence to structure match of regions with low sequence similarity, where the alignment is error prone between correctly predicted secondary structure element of query sequence and the corresponding native secondary structure string of template structure.

In addition to SA, another important geometrical parameter to characterize the effect of solvation is residue depth (RD). RD of an amino acid residue in a protein is defined as the average distance of its atoms from the nearest surface water molecule.[54] RD is originally developed to study protein interior (e.g., to distinguish residues close to the protein surface and those deep within the protein interior) and is shown to correlate better than SA with protein stability, because it provides a more detailed description of residue burial.[55] The application of RD information to fold recognition, unlike SA, is introduced only recently in a method[56] called SP^3. In SP^3, RD is used to derive a depth-dependent, structure-based sequence profile of template. The depth-dependent sequence profile is, then, combined with secondary-structure and sequence-derived profiles for fold recognition. SP^3 is one of the best performing single-method servers in CASP 6.[57,58]

Although both SA and RD describe the level of solvent exposure of a residue, studies[55,59,60] indicate that they provide substantially different, complementary information and there is no simple correlation between them. This raises an interesting question: can SA be incorporated into SP^3 to further improve the accuracy of fold recognition?

In this work, we address this question by developing a fold recognition method, called SP^4. SP^4 integrates sequence-derived profile, secondary structure profile, RD-dependent structure-based profile, and solvent accessibility profile to recognize structural homologs. To our knowledge, none of the previously published work has incorporated both RD and SA information for fold rec-
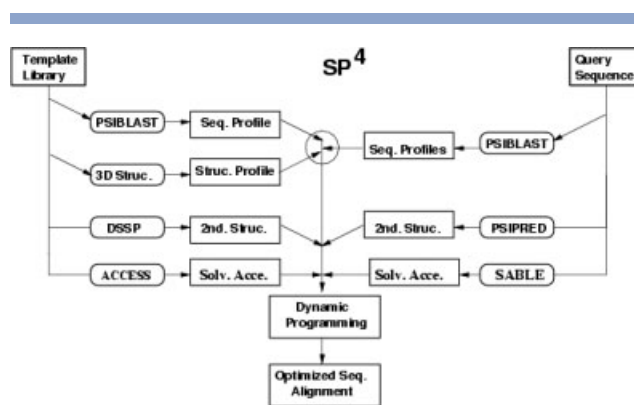

Figure 1
The flow chart of SP^4 for query-template alignment.

ognition. We show that SP^4 provides a robust improvement in remote-homolog recognition comparing with its variants lacking either RD or SA or both.

METHOD

The SP^4 method is built on the SP^3 method[56] by incorporating the relative solvent accessibility profile. The flowchart of SP^4 algorithm for pairwise query-template alignment is shown in Figure 1. The details are as follows.

Profile generation

For matching a query sequence to a template structure, the SP^4 algorithm requires four profiles as input. They are (1) an RD-dependent sequence profile derived from the template structure, (2) sequence profiles generated from query and template sequences, (3) predicted and actual secondary structures of the query sequence and the template structure, respectively, and (4) predicted and actual SA of the query sequence and the template structure, respectively.

RD-dependent structure-derived sequence profile

A template structure is divided into structural fragments with a sliding window of 9 amino acid residues along the template sequence. Each structural fragment of the template is compared to same-size fragments in a collection of 1011 high-resolution, nonhomologous protein structures.[61] The similarity score between a template fragment and a fragment in the structural database is given by the equation

$$S_{str} = d_{rmsd}^2 + 10 \times \sum_{j \in [1,9]} [\exp(-D_j^{template}/2.8)$$
$$- \exp(-D_j^{database}/2.8)]^2, \quad (1)$$

where d_{rmsd} is the root-mean-squared distance (RMSD) between the two fragment structures, $D_j^{template}$ and $D_j^{database}$

are the depth[55] 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. The depth (in Å) is scaled by 2.8 Å, the diameter of a water molecule. For each template fragment, the sequences of top 25 fragments ranked by the similarity score S_{str} are retained for constructing the structure-derived sequence profile.

Sequence profiles

The sequence profiles are constructed by three iterations of PSIBLAST[5] searching (E value cutoff of 0.001) against nonredundant (NR) sequence database, which was filtered to remove low-complexity regions, transmembrane regions, and coiled-coil segments.[62] Sequence profiles for the sequences of both query and template are obtained.

Secondary-structure

The secondary structure of query sequence with a three-state classification (helix, strand, and coil) is predicted by PSIPRED.[62] The secondary structures of templates are obtained by MolAuto[63] using a DSSP-like algorithm.[64]

Solvent accessibility

The solvent accessibility of query sequence with a two-state classification (buried and exposed based on a 25% SA threshold) is predicted by SABLE .[65] The residue SAs of templates are obtained by the ACCESS algorithm[42] (normalized by the extended state solvent-accessible surface areas which are obtained from Chothia[66]).

Profile–profile comparison

The local–local dynamic programming method[67] is used to optimize the score that matches the query profiles with template profiles. The alignment score for aligning query sequence position i with the template sequence position j is given by the equation

$$
\begin{aligned}
S(i,j) = &-(1 - w_{struc})F_{query}^{seq}(i) \cdot M_{template}^{seq}(j) \\
&- w_{struc}F_{template}^{struc}(j) \cdot M_{query}^{seq}(i) \\
&- w_{2ndary}\delta_{si,sj} - w_{sa}\delta_{ai,aj} + s_{shift}, \quad (2)
\end{aligned}
$$

where $F_{query}^{seq}(i)$ is the sequence-derived frequency profile of the query sequence, $M_{template}^{seq}(j)$ is the sequence-derived log odd profile (position-specific substitution matrix as in PSIPRED) of the template, $F_{template}^{struc}(j)$ is the RD-dependent structure-derived frequency profile of the template, $M_{query}^{seq}(i)$ is the sequence-derived, log odd profile of the query sequence, s_{shift} is a to-be-determined constant shift, w_{struc}, w_{2ndary}, and w_{sa} are the three weight parameters for RD-dependent structure-derived sequence

profiles, secondary structure profiles, and solvent accessibility profiles, respectively, and $\delta_{si,sj}$ (or $\delta_{ai,aj}$) 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 (or ri and rj, the states based on the SA):

$$
\delta_{m,n} = \begin{cases} 1 & m = n, \\ -1 & m \neq n, \end{cases}
$$

where $m = si$ and $n = sj$, or $m = ri$ and $n = rj$. A secondary-structure dependent gap penalty is employed. That is, no gaps are allowed if $si = sj = \alpha$ (helix) or $si = sj = \beta$ (sheet). The gap opening (w_0) and gap extension (w_1) penalties are applied to other regions.

Template ranking

We used the same empirical method as in SP[3] for template ranking. Briefly, the templates are ranked based on the difference between the raw alignment score and the reverse alignment raw score in which the alignment is made with the reversed query sequence.[6] If there is no structural similarity between first two models (defined as zero MaxSub score[68]), templates will be re-ranked by the larger one of the two Z-scores, which are calculated based on the raw alignment score normalized by the full alignment length and the non-end-gap alignment length, respectively. Here, the Z-score for a template i is given by $Z(i) = [S_n(i) - S_n^{ave}]/S_n^{sd}$, where ave and sd denote the average and standard deviation of normalized scores for all the templates.

Method variants, parameterization and Testing

For a better understanding of the role of SA and RD in the accuracy of fold recognition, we compare many variants of SP methods that include SP[1] (Sequence profile only), SP[2] (Sequence profile + secondary structures), SP[3] (Sequence profile + secondary structures + RD-dependent structure-derived sequence profile), SP[2+] (Sequence profile + secondary structures + SA), and SP[4] (Sequence profile + secondary structures + RD-dependent structure-derived sequence profile + SA). There are six adjustable parameters (w_0, w_1, w_{2ndary}, w_{sa}, w_{struc}, and s_{shift}) for SP[4] and five for SP[2+]. The parameters for all methods are optimized separately for the best alignment accuracy for the ProSup benchmark.[69] The parameters for SP, SP[2], and SP[3] were obtained previously.[56] The parameters for SP[2+] and SP[4] are obtained in this work.

All SP methods are tested in SALIGN,[11] Lindahl,[70] and LiveBench 8[71] benchmarks. To facilitate comparison with early SP methods, we use the original profiles[56] used in SP[1], SP[2], and SP[3] for ProSup, Lindahl, and LiveBench 8 benchmarks. In the SALIGN benchmark, new profiles are generated for all SP methods. However, it

Table I

The Alignment Accuracy for ProSup Training Set

Method	Accuracy^a (%)	±4 Residues^b (%)
SP^1	55.9	72.4
SP^2	62.9	79.2
SP^{2+}	64.4	80.5
SP^3	65.3	82.2
SP^4	66.8 (68.1)^c	83.8

^{a}One-to-one match given by the method and benchmark.

^{b}Match within 4 residues from one-to-one match.

^{c}The number in parenthesis is the alignment accuracy if the exposed and buried states are obtained from the native structure of the query sequence rather than from prediction.

should be emphasized that the comparison made between SP methods and other published methods is not a strict one, because of the limitation of time-dependent library used by different methods. Both SP^3 and SP^4 are also evaluated in the recently completed CASP7 blind test.

RESULTS

Training set: ProSup benchmark

The ProSup benchmark consists of 127 pairwise protein alignments obtained by structural alignment program ProSup.[69] We optimized the parameters of SP^{2+} and SP^4 by maximizing the percent of matches between the structural alignment and the alignment made by the SP^x fold-recognition method. The optimization is done by sequential grid-search until further iterations do not improve the alignment accuracy.[56] The final optimized parameter sets are shown in the supplement material (Table S1).

Table I compares the performance of SP methods. After optimization, introducing the SA term to SP^2 and SP^3 makes 1.5% and 1.5% in alignment accuracy, respectively. Similar improvements are observed if we define correct alignment as a match within four residues from structural alignment. This can be compared to 7, 2.4, and 1% improvements from adding secondary structures (SP^1 to SP^2), RD-dependent, structure-derived sequence profiles (SP^2 to SP^3), and RD-independent structure-derived sequence profiles, respectively.

SP^4 achieves the highest alignment accuracy (66.8%). This is 1.5% higher than that from SP^3 (without SA), 2.4% than from SP^{2+} (without RD dependent term), and 3.9% than from SP^2 (without both RD and SA). The trend of additive improvement is consistent with the complementary nature of RD and SA. Table I also shows that if the exact SA states (rather than predicted) are used in SP^4, the alignment accuracy can be further improved by another 1%.

Test set 1: SALIGN benchmark

To test the alignment accuracy, we used the SALIGN benchmark.[11] This benchmark contains 200 selected pairs with an average pair sharing 20% sequence identity and 65% of structurally equivalent C$_\alpha$ atoms superposed with an rmsd of 3.5 Å.[11] We use SALIGN benchmark as an independent test from the ProSup benchmark for alignment accuracy because sequence identities between any pairs of proteins from the two benchmarks are less than 40%. Alignment accuracy for this benchmark is assessed by calculating the fraction of the alignment that is the same as the structural alignment obtained from the TMalign program[72] [i.e., TM overlap]. We use TMalign instead of CE program[73] because TMalign is a global alignment program which yields a single alignment between the whole protein pairs by optimizing their TMscore. In contrast, CE is a local alignment program which often produces several alignments between different (sometimes overlapped) fragment pairs conferring structural similarity. These alternative alignments between overlapped fragment pairs prevent a straightforward evaluation of different SP methods.

The alignment accuracies of different SP methods in the SALIGN benchmark are shown in Table II. Introduction of the SA term (SP^2 to SP^{2+} and SP^3 to SP^4) leads to at least 1% consistent improvement in alignment accuracy. This improvement is smaller than that from the use of secondary structure (2.7% from SP^1 to SP^2) or the use of RD-dependent structure-derived sequence profiles (2.1% from SP^2 to SP^3). Nevertheless, we observed here again that concurrent use of RD and SA made an additive improvement. That is, the 3.1% improvement of SP^4 over SP^2 roughly equals to the sum of improvement of SP^4 over SP^3 (1.0%) and that of SP^4 over SP^{2+} (1.8%). Similar results are obtained if the assessment of alignment accuracy is made only on those most reliably aligned residues. These reliably aligned residues (i.e., TM core overlap[72]) are the pair of residues whose C$_\alpha$ atom distance between the query structure and template structure is less than 5 Å. To make clearer the statistical significance of the difference among SP methods, Figure 2 shows the alignment accuracies of

Table II

The Alignment and Model Accuracy for the SALIGN Benchmark

Method	SP^1	SP^2	SP^{2+}	SP^3	SP^4
TM overlap^a (%)	51.5	54.2	55.5	56.3	57.3
Core TM overlap^b (%)	53.8	56.6	58.0	58.8	59.9
Total MaxSub Score^c	68.7	70.3	71.8	74.1	75.3

^{a}One-to-one match given by the method and TMalign.

^{b}One-to-one match, only the residue pairs of distance <5 Å between query and template are counted.

^{c}Total MaxSub score for the models built based on the fold-recognition alignment.
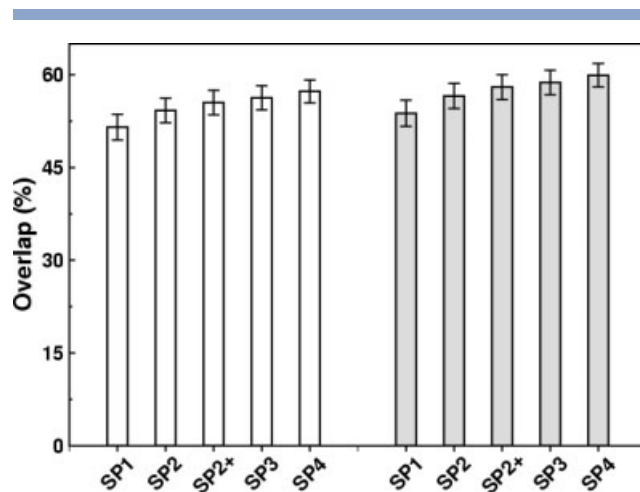
Figure 2

The alignment accuracies of different SP methods in SALIGN benchmark. The open bars on the left and gray bars on the right are for TM overlaps and TM core overlaps, respectively. (See the Results section.)

various methods along with error bars of one standard deviation of the mean.

Another way to measure the performance of SP^4 is to calculate the accuracy of the model built from the fold-recognition alignment. This is done by transferring the C_α coordinates of the template structures to the aligned residues in the query sequence. The accuracy of the resulting model can be assessed by the MaxSub score between the model and the known native structure. MaxSub score[68] 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). Table II reports the sum of the MaxSub scores for the models built by various SP methods in the SALIGN benchmark. The magnitude of improvement in MaxSub scores and the additive trend is similar to that in alignment accuracy. This indicates the improvement of SP^4 over SP^3 and SP^4 over SP^{2+} not only in alignment but also in the structural models produced.

The improvement of alignment accuracy due to the SA and RD terms in SALIGN benchmark, although is slightly smaller than that in the training ProSup benchmark, is remarkable, considering that all SP methods are trained by a different structural alignment program called ProSup whereas the SALIGN benchmark is evaluated according to the TM structural alignment. It is known that different structural alignment program might generate different alignment for structurally variable regions.[72,74] Thus, the improvement of alignment accuracy due to the SA and/or RD terms is independent

of the method that is used for measuring alignment accuracy.

Test set 2: Lindahl benchmark

The Lindahl set contains 976 proteins, with 555, 434, and 321 pairs of proteins in the same family, superfamily, and fold, respectively.[70] The fold-recognition sensitivity of each method is tested by aligning each protein with the rest 975 proteins, and checking whether or not the method can recognize the member of same family, superfamily or fold as the first rank or within the top 5 ranks. Results obtained from various SP methods and several other methods are listed in Table III. We use the original profiles from Ref. 56 to facilitate the comparison within SP methods (see Method section). We emphasize that the comparisons between SP methods and other published methods only serve as an approximate guide because of the time dependent nature of sequence database for sequence profiles.

Within SP methods, the concurrent use of RD and SA consistently improves the success rates for recognizing proteins within superfamily or fold. The biggest improvement of SP^4 over SP^3 is 6.2% on ranking top 5 at the fold level. There is a slight reduction at the family level (-0.7% from SP^3 to SP^4). This suggests that the most significant improvement of SP^4 over SP^3 is on recognizing remote homologs. Indeed, the best success rate within SP methods on ranking top 5 at the family level is

Table III

The Success Rates for Recognizing Proteins Within the Same Family, Superfamily, or Fold in the Lindahl Benchmark

Method	Family only (%)		Superfamily only (%)		Fold only (%)	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
PSIBLAST[a]	71.2[b]	72.3	27.4	27.9	4.0	4.7
HMMER[a]	67.7	73.5	20.7	31.3	4.4	14.6
SAMT98[a]	70.1	75.4	28.3	38.9	3.4	18.7
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
SPARKS-0[e]	82.9	**90.1**	56.5	**72.4**	23.1	43.6
FOLDpro[f]	**85.0**	89.9	55.5	70.0	26.5	48.3
SP[1]	81.3	87.2	51.8	65.0	20.2	35.8
SP[2]	82.5	87.0	52.5	67.1	24.9	40.2
SP[2+]	82.2	86.8	55.1	68.4	27.8	45.8
SP[3]	81.6	86.8	55.3	67.7	28.7	47.4
SP[4]	80.9	86.3	**57.8**	68.9	**30.8**	**53.6**

[a]From Ref. 21.
[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 5 rank of the template. The highest result within a category is highlighted in bold.
[c]From Ref. 22.
[d]From Ref. 24.
[e]From Ref. 44. SPARKS-0 is SPARKS without the structure-derived score.
[f]Results from 10-fold cross validation, Ref. 27.

Table IV
The Model Quality for Lindahl Benchmark

Method	SP^1	SP^2	SP^{2+}	SP^3	SP^4
Total[a]	328.6[b]	340.8	343.4	349.2	352.5
Family[c]	286.8	292.8	292.7	293.5	295.6
Superfamily[d]	87.5	94.3	99.9	100.8	108.9
Fold[e]	21.7	27.8	30.2	34.5	37.1

[a]All 976 proteins.
[b]The summed MaxSub score for the first-ranked models.
[c]Family only.
[d]Superfamily only.
[e]Fold only.

achieved by SP^1 which only includes sequence-derived profiles. Among all methods listed, SP^4 has the highest success rate on the fold level (both first and top 5 ranks) and the superfamily level for the first rank.

The above results are based on somewhat subjective SCOP classification,[75] which may not reflect the true accuracy. To further confirm the improvement due to SA and RD, we calculated the MaxSub score for the structural models built from their respective first-ranking templates. As shown in Table IV, the use of either SA or RD dependent term improves MaxSub scores at almost all levels. For SA term (i.e., from SP^2 to SP^{2+}), the improvement is not visible at the family level (0%) and more significant at more difficult superfamily (6%) and fold (9%) levels. For RD dependent term (i.e., from SP^2 to SP^3), the improvement is also small at the family level and more significant at the superfamily (7%) and fold (24%) levels. The use of both SA and RD dependent term (i.e. SP^4) is at least 0.7, 8.0, and 7.5% better than either SP^3 or SP^{2+} at the family, superfamily, and fold level, respectively. This further suggests that the information of solvation is more useful for improving remote-homolog recognition when the sequence identity between query and template is low. For the whole benchmark, the improvement of SP^4 over SP^2 (11.7) in raw MaxSub

score is similar to the summed improvement of SP^4 over SP^{2+} (9.1) and SP^4 over SP^3 (3.3).

Test set 3: LiveBench 8 benchmark

LiveBench 8 consists of 172 targets which could be further divided into 99 "hard" and 73 "easy" ones based on whether a target could match to known PDB structures with PSI-BLAST *e*-value < 0.01.[71] The comparison between different methods is based on sensitivity, specificity, and total MaxSub scores (accuracy). 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. Here, all SP results are based on C_α models of aligned residues.

As shown in Table V, the improvement of SP^4 over SP^3 (or SP^{2+}) is mostly on "hard" targets. SP^4 is 9.6 and 21.1% better than SP^3 in terms of sensitivity and specificity, respectively. SP^4 is also 7.5 and 11.5% better than SP^{2+} on sensitivity and total MaxSub scores. Comparing with SP^2 which does not count either RD or SA information, SP^4 is 14, 19, and 15% better in sensitivity, specificity, and MaxSub scores, respectively.

Test set 4: CASP7 blind test set

The recently completed CASP7 consists of 95 targets and was carried out between May and July of 2006, with the participation of about 68 single/meta servers and 119 expert groups (for tertiary structure predictions category). The 95 targets are officially classified into 108 template-based-modeling (TBM) domains and 19 free-modeling (FM) domains, based on whether or not the structurally similar template (deposited in PDB) had been identified and used in prediction. The performance of different methods are officially evaluated based on the GDT Z score[76] calculated by official CASP7 assessors (http://www.predictioncenter.org/casp7/).

Like SPARKS2 and SP^3, SP^4 was directly involved in CASP7 blind test as an automatic server. The template

Table V
The Performance of SP Methods in LiveBench 8 Benchmark

Method	Hard targets only (99)[a]			Easy targets only (73)[a]		
	Sens.[b]	Spec.[c]	MaxSub[d]	Sens.[b]	Spec.[c]	MaxSub[d]
SP^1	39	23.7	9.65	67	67.0	27.53
SP^2	50	30.4	11.46	66	65.7	28.05
SP^{2+}	53	35.8	11.85	66	65.1	28.11
SP^3	52	29.9	13.20	68	67.4	29.04
SP^4	57	36.2	13.21	69	66.8	28.95

[a]Based on whether a target could match a known PDB structure with PSI-BLAST e-value less than 0.01 (http://BioInfo.PL).
[b]Sensitivity.
[c]Specificity.
[d]Total MaxSub scores.

Table VI
The Performance in CASP7 Blind Prediction[a]

Method	Top 1			Top 5		
	FM[b]	TBM[c]	ALL[d]	FM[b]	TBM[c]	ALL[b]
SPARKS2	−2.1	35.5	36.1	1.5	42.8	44.0
SP^3	2.7	44.8	47.4	7.5	48.7	55.1
SP^4	3.3	45.2	47.9	12.8	55.1	66.6

[a]The GDT-Z score is based on the official CASP7 assessment.
[b]19 Free modeling targets.
[c]108 Template-Based Modeling targets.
[d]All 124 targets. (There are four targets belonging to TBM/FM according to official CASP7 assessors.)
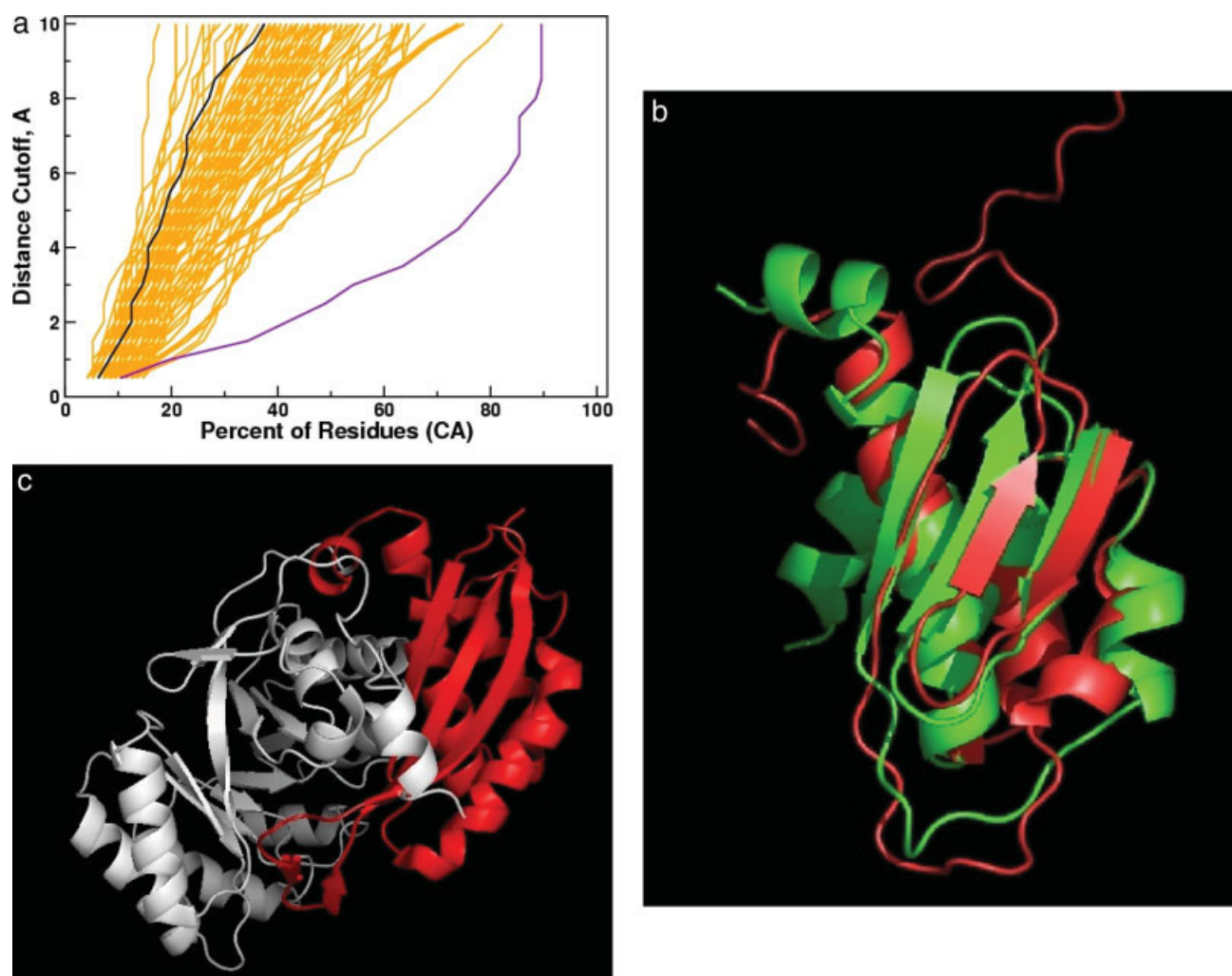
Figure 3

(**a**) *GDT plot of T0321_D1 for top-1 model. The plot was generated from the official CASP7 assessment (http://www.predictioncenter.org/casp7/). The x-axis is the largest subset of residues in the model that can be fitted to the target in a rigid body sequence dependent superposition, and the y-axis is the cutoffs ranging from 0.5 to 10.0 Å. Results are plotted as a line for each model separately, with the models from SP^4, SP^3 and all other methods (server and expert) colored violet, black, and orange, separately.*
(**b**) *The superposition between the native structure of T0321_D1 (green, pdb code: 2h1q) and the top-1 model (red) generated with modeller based on SP^4 prediction.*
(**c**) *The template (pdb code: 2mnr_) identified by SP^4 server, with the region aligned to top-1 model of SP^4 colored by red.*

library for SP^4, SP^3, and SPARKS2 was built in the same way. 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 1.61 if they have less than 40% sequence identity with the sequences already in the library.

As shown in Table VI, the improvement of SP^4 over SP^3 is more evident on the level of Top-5 model comparing with that on the level of Top-1 model, and the difference is more significant on the FM targets comparing with that on the TBM ones. For all the 124 domains, the GDT Z score of Top-5 model by SP^4 is 20% higher than that of SP^3 and ranked No. 5 among all 68 CASP7 serv-

ers (following Zhang-Server, ROBETTA, Pmodeller6, and RAPTOR). Based on the independent visual assessment of FM targets by official assessors, SP^4 is also ranked No. 5 among all servers by following Zhang-Server, ROBETTA, Pmodeller6, and Keasar-server (http://www.predictioncenter.org/casp7/, free modeling presentation by Neil Clarke). Unlike SP^4, many of the top servers are belong to consensus methods.

Here, we present the prediction of T0321_D1 (target 321, domain 1) by SP^4 as a successful example. Figure 3 (a) shows that the quality of top-1 model by SP^4 is significantly better than those predicted by all other expert/server methods including SP^3. Beside the top 1 model, there are three other ones within the top 5 models by SP^4 which are also

among the best models available (not shown). In the top-1 model, 74 of 96 residues (77%) can be superposed to the native structure based on LGA tool[76] with a 5 Å cutoff, and the fit RMSD is 2.9 Å [Fig. 3(b)]. SP^4 produced this model because it successfully recognized the structurally similar fragment from a large protein [Fig. 3(c)]. The remarkably different accuracy between SP^4 and SP^3 for this target indicates the power of concurrent use of SA and RD in fold recognition.

DISCUSSION

In this article, the method SP^4 is developed by incorporating predicted and actual solvent-accessibility profiles into SP^3. Testing on SALIGN, Lindahl, LiveBench 8 benchmarks, and CASP7 blind prediction all indicates that the new method improves over SP^3 not only in the accuracy of sequence alignments and predicted model structures but also in the sensitivity of detecting remote homolog with the same structural fold. In SP^3, the solvent exposure of residues has been somewhat taken into account through the use of RD in generating structure-derived profiles. An improvement of SP^4 over SP^3 (and SP^{2+}) confirms that SA and RD are complementary in describing the effect of solvation.[55,59,60] More importantly, the trend of additive improvement we observed suggest that SP^4, by concurrent use of SA and RD, has captured their complementary information to improve the accuracy and sensitivity of fold-recognition.

In this assessment of the usefulness of SA in improving SP^3, both predicted and actual SA profiles are based on two states (exposed and buried) classified according to an arbitrary threshold of 25%. The two-state accuracy by SABLE[65] is 77.3% in the ProSup benchmark, 77.9% in the SALIGN benchmark, 74.3% in the Lindahl benchmark and, 75.3% in the LiveBench 8 benchmark. (The accuracy based on Matthews correlation coefficient is also shown in Supplement material, Table S2.) This accuracy is consistent with the published performance of this and other state-of-the-art predictors.[46,77,78] A recent method SPINE reaches a 10-fold cross-validated accuracy of 79% for a large set of 2640 proteins.[79] We found that SP^4 based on actual SA profiles for both query and templates improve the alignment accuracy by an additional 1% for ProSup. This indicates that the performance of SP^4 can be further improved with a more accurate method for SA prediction.

The two-state classification increases the accuracy of prediction by reducing number of states in SA. This is at the cost of losing the detailed fluctuation pattern of SA along the sequence. As shown in Figure S1, residues with a certain RD value could span a wide range of SA value,[55,60] and an arbitrary cutoff of SA (e.g., 25% used in this work) could only capture a snapshot of the "local fingerprints" of 3D packing. Thus, it remains to be tested

whether a SA profile based on the multiple classes or real value prediction,[46,80–82] when coupled with RD information, can further improve the accuracy of fold-recognition. The work is in progress.

One interesting question is whether or not the information derived from the three-dimensional structure of template is useful for improving fold-recognition. It was found difficult to harness three-dimensional structural information in fold-recognition.[83] The observation that SP^4 is more accurate than either SP^{2+} or SP^3 indicates the importance of integrating sequence, 2D (SS and SA), and 3D derived information together for detection of remote homologs.

One extension of this work is to replace SP^2 used in multiple sequence alignment method SPEM[84] by SP^{2+}. SPEM combines the pairwise alignment generated from SP^2 with a progressive algorithm to generate multiple sequence alignment. The method provides significant improvements in aligning remote homologs when comparing with the state-of-the-art techniques such as ClustalW.[85] Because SP^{2+} consistently improves over SP^2 in all benchmarks tested here, we expect that it will further increase the quality of multiple sequence alignment (for the multiple alignment of remote homologs, in particular).

REFERENCES

1. Goldsmith-Fischman S, Honig B. Structural genomics: computational methods for structure analysis. Protein Sci. 2003;12:1813–1821.
2. Friedberg I, Jaroszewski L, Ye Y, Godzik A. The interplay of fold recognition and experimental structure determination in structural genomics. Curr Opin Struct Biol 2004;14:307–312.
3. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003;334:793–802.
4. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and completeness of single domain structures. Proc Natl Acad Sci USA 2006;103:2605–2610.
5. 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.
6. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
7. Koretke KK, Russell RB, Lupas AN. Fold recognition from sequence comparisons. Proteins 2001;5(Suppl):68–75.
8. 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.
9. 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.
10. Pei J, Sadreyev R, Grishin N. PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 2003;19:427–428.
11. Marti-Renom MA, Madhusudhan M, Sali A. Alignment of protein sequences by their profiles. Protein Sci 2004;13:1071–1087.
12. Edgar RC, Sjolander K. COACH: profile-profile alignment of protein families using hidden Markov models. Bioinformatics 2004;20:1309–1318.
13. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.

14. Wistrand M, Sonnhammer E. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. BMC Bioinformatics 2005;6:99.

15. Tan YH, Huang H, Kihara D. Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. Proteins 2006;64:587–600.

16. Elofsson A, Fischer D, Rice DW, Le Grand SM, Eisenberg D. A study of combined structure/sequence profiles. Fold Des 1996;1:451–461.

17. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.

18. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.

19. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.

20. 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.

21. 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.

22. Xu J, Li M, Lin G, Kim D, Xu Y. Protein structure prediction by linear programming. Pac Symp Biocomput 2003;8:264–275.

23. 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.

24. 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.

25. Torda AE, Procter JB, Huber T. Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. Nucleic Acids Res 2004;32:W532–W535.

26. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004;56:502–518.

27. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463.

28. 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.

29. Godzik A. Fold recognition methods. Methods Biochem Anal 2003;44:525–546.

30. Kinch L, Wrabl J, Krishna S, Majumdar I, Sadreyev R, Qi Y, Pei J, Cheng H, Grishin N. CASP5 assessment of fold recognition target predictions. Proteins 2003;6 (Suppl):395–409.

31. Wang G, Dunbrack RL, Jr. Scoring profile-profile sequence alignments. Protein Sci 2004;13:1612–1626.

32. Edgar RC, Sjolander K. A comparison of scoring functions for protein sequence profile alignment. Bioinformatics 2004;20:1301–1308.

33. Petrey D, Honig B. Protein structure prediction: inroads to biology. Mol Cell 2005;20:811–819.

34. Dunbrack RL, Jr. Sequence comparison and protein structure prediction. Curr Opin Struct Biol 2006;16:374–384.

35. Bujnicki J, Elofsson A, Fischer D, Rychlewski L. Structure prediction meta server. Bioinformatics 2001;21:750–751.

36. Juan D, Grana O, Pazos F, Fariselli P, Casadio R, Valencia A. A neural network approach to evaluate fold recognition results. Proteins 2003;50:600–608.

37. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;22:1015–1018.

38. Chivian D, Kim D, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss C, Bonneau R, Rohl AC, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53:524–533.

39. Wallner B, Fang H, Elofsson A. Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. Proteins 2003;53:534–541.

40. Fischer D. Servers for protein structure prediction. Curr Opin Struct Biol 2006;16:178–182.

41. Valencia A. Meta, Meta^n and Cyber servers. Bioinformatics 2003;19:795–795.

42. Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.

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

44. 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.

45. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.

46. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753–767.

47. Rost B, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;270:471–480.

48. McGuffin LJ, Jones DT. Benchmarking secondary structure prediction for fold recognition. Proteins 2003;52:166–175.

49. 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.

50. Rost B. Topits: threading one-dimensional predictions into three-dimensional structures. In: Rawlings C, Clark D, Altman R, Hunter L, Lengauer T, Wodak S, editors. Third International Conference on Intelligent Systems for Molecular Biology. Cambridge, England: AAAI Press; 1995. pp 314–321.

51. Karchin R, Cline M, Karplus K. Evaluation of local structure alphabets based on residue burial. Proteins 2004;55:508–518.

52. Przybylski D, Rost B. Improving fold recognition without folds. J Mol Biol 2004;341:255–269.

53. Qiu J, Elber R. SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. Proteins 2006;62:881–891.

54. Pedersen T, Sigurskjold B, Andersen K, Kjaer M, Poulsen F, Dobson C, Redfield C. A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution. J Mol Biol 1991;218:413–426.

55. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Struct Fold Des 1999;15:723–732.

56. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328.

57. Zhou H, Zhou Y. SPARKS 2 and SP^3 servers in CASP6. Proteins 2005;7 (Suppl):152–156.

58. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;7 (Suppl):27–45.

59. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophys J 2003;84:2553–2561.

60. Pintar A, Carugo O, Pongor S. Atom depth in protein structure and function. Trends Biochem Sci 2003;28:593–597.

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. Kraulis P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24:946–950.

64. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.

65. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59: 467–475.

66. Chothia C. The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–12.

67. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.

68. 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.

69. 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.

70. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. J Mol Biol 2000;295:613–625.

71. Rychlewski L, Fischer D. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 2005;14:240–245.

72. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.

73. Shindyalov IN, Bourne P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.

74. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005;346:1173–1188.

75. 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.

76. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.

77. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.

78. Chen HL, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 2005;33:3193–3199.

79. Dor O, Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 2007;66:838–845.

80. Garg A, Kaur H, Raghava G. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 2005;61:318–324.

81. Xu Z, Zhang C, Liu S, Zhou Y. QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. Proteins 2006;63:961–966.

82. Dor O, Zhou Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins, in press.

83. 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.

84. Zhou H, Zhou Y. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 2006;21:3615–3621.

85. Thompson J, Higgins D, Gibson T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4690.