SP\(^5\): Improving Protein Fold Recognition by Using Torsion Angle Profiles and Profile-Based Gap Penalty Model

Wei Zhang\(^1,2\), Song Liu\(^3,4,5\), Yaoqi Zhou\(^1\)*

\(^1\) Indiana University School of Informatics and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana, United States of America, \(^2\) Institute of Applied Physics and Computational Mathematics, Beijing, People’s Republic of China, \(^3\) Department of Biostatistics, Center of Excellence in Bioinformatics & Life Sciences, University at Buffalo, State University of New York, Buffalo, New York, United States of America, \(^4\) Department of Biostatistics, Roswell Park Cancer Institute, Buffalo, New York, United States of America, \(^5\) Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology & Biophysics, University at Buffalo, State University of New York, Buffalo, New York, United States of America

Abstract

How to recognize the structural fold of a protein is one of the challenges in protein structure prediction. We have developed a series of single (non-consensus) methods (SPARKS, SP\(^2\), SP\(^3\), SP\(^5\)) that are based on weighted matching of two to four sequence and structure-based profiles. There is a robust improvement of the accuracy and sensitivity of fold recognition as the number of matching profiles increases. Here, we introduce a new profile-profile comparison term based on real-value dihedral torsion angles. Together with updated real-value solvent accessibility profile and a new variable gap-penalty model based on fractional power of insertion/deletion profiles, the new method (SP\(^5\)) leads to a robust improvement over previous SP method. There is a 2% absolute increase (5% relative improvement) in alignment accuracy over SP\(^2\) based on two independent benchmarks. Moreover, SP\(^5\) makes 7% absolute increase (22% relative improvement) in success rate of recognizing correct structural folds, and 32% relative improvement in model accuracy of models within the same fold in Lindahl benchmark. In addition, modeling accuracy of top-1 ranked models is improved by 12% over SP\(^4\) for the difficult targets in CASP 7 test set. These results highlight the importance of harnessing predicted structural properties in challenging remote-homolog recognition. The SP\(^5\) server is available at http://sparks.informatics.iupui.edu.

Introduction

Fold recognition refers to recognizing the structural fold of a protein, given its sequence information. Fold recognition is one of the key bottlenecks for protein structure predictions as the protein data bank now appears to contain the complete (or near complete) set for all possible structural folds of proteins, at least for small domain proteins [1,2].

Recently completed assessment of automated servers for protein structure prediction (CASP 7) [3] reveals the power of post-treatment of models predicted by individual fold recognition methods through consensus predictions (For example, ROBETTA [4], Pmodeller [5], Fams-ace [6] and/or constrained template-fragment recombination and refinement (For example, Chunk-TASSER [7], I-TASSER [8]). The prediction quality of these methods, however, relies heavily on the accuracy of initial models generated by individual fold recognition methods in the first step. Another observation is that the accuracy of top single servers can rival with most consensus methods. Thus, developing and/or improving individual methods are critically important for further advancement of the accuracy of fold recognition and structure prediction.

We have developed a series of single fold-recognition methods (SPARKS, SP\(^2\), SP\(^3\), SP\(^5\)) that are based on weighted matching of multiple profiles that include sequence profiles generated from multiple sequence alignment [9], predicted versus actual secondary structures [10,11], knowledge-based profile (single-body) score function [10], depth-dependent sequence profiles derived from template structures [11], and predicted versus actual solvent accessible surface area [12]. There is a robust improvement of the accuracy and sensitivity of fold recognition as the number of matching profiles increases [10,11], and [12]. SPARKS, SP\(^3\), and SP\(^5\) were ranked among the top performers for automatic servers in recent CASP 6 [13,14] and 7 [12,3]. This exemplifies the importance and effectiveness of multiple-dimensional use of the structural information of templates in developing fold-recognition techniques.

In this paper, we introduce the fifth “dimension” for fold recognition by incorporating predicted backbone torsion angles (SP\(^5\)). The backbone torsion angles (\(\phi\) and \(\psi\)) are two rotation angles about the \(\text{C}_\alpha – \text{N}\) bond (\(\phi\)) and the \(\text{C}_\alpha – \text{C}\) bond (\(\psi\)). Because the polypeptide backbone of a protein is a linked sequence of rigid planar peptide groups, these two angles essentially determine the backbone conformation of proteins. While a three-state classification of secondary structures is a coarse-grained one-dimensional representation of local backbone conformation, backbone torsion angles encode the backbone tertiary structure, at least in principle.
Traditionally, dihedral torsion angles are predicted as a few discrete states based on local (fragment) structural patterns using either machine-learning techniques or classification schemes [15–22]. However, there were only a few limited applications of predicted angle states to fold recognition [19] and sequence alignment [23]. The former uses torsion-angle states as a replacement of simple three-state secondary structures to build an iterated alignment hidden Markov model [18]. The latter [23] predicts angle states by hidden Markov model and employs the predicted angles to build structural context-based substitution matrices. Here, we propose to match predicted and actual torsion angles as a new profile term in a multi-dimensional profile-profile alignment. This represents a novel use of predicted torsion angles as a complementary to rather than a replacement of secondary structures for fold recognition. The angle profile used in this work is built on a recent advancement in real-value prediction of torsion structures for fold recognition. The angel profile used in this work is found comparable to or more accurate than those based on multi-state classification of the $\phi - \psi$ map.

In SP$^3$, the effect of solvation was taken into consideration by matching the predicted and actual solvent accessibility (SA). The SA profiles are based on two states (exposed and buried) classified according to an arbitrary threshold of 25%. 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. We recently have developed method (called Real-SPINE) for real value SA prediction, which yields a 10-fold cross-validated mean absolute errors of 38% for $\psi$ and 25% for $\phi$ [24]. This accuracy of real-value prediction was found comparable to or more accurate than those based on multi-state classification of the $\phi - \psi$ map.

Table 1. The alignment accuracies for Prosop and SALIGN benchmark.

<table>
<thead>
<tr>
<th></th>
<th>SP$^3$</th>
<th>SP$^4$</th>
<th>SP$^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosop</td>
<td>65.3 $\pm$ 0.22%</td>
<td>66.8 $\pm$ 0.20%</td>
<td>68.7 $\pm$ 0.20%</td>
</tr>
<tr>
<td>SALIGN</td>
<td>56.3 $\pm$ 0.14%</td>
<td>57.3 $\pm$ 0.13%</td>
<td>59.7 $\pm$ 0.15%</td>
</tr>
</tbody>
</table>

*One-to-one match given by the method and Prosop.
*One-to-one match given by the method and TMalign.

The alignment accuracy of the methods trained by PREFAB benchmarks is tested by the Prosop and SALIGN benchmarks. Prosop benchmark, prepared by Sipp’s group, consists of 127 pairs of proteins with alignment by structural alignment program Prosop [38]. SALIGN benchmark [42] contains 200 selected pairs with an average pair sharing 20% sequence identity and 63% of structurally equivalent $C_\alpha$ atoms superposed with an rmsd of 3.5 Å [42]. Reference alignment is obtained from the structural alignment obtained from the TMalign program [43] [i.e., TM overlap]. The sequence identity between PREFAB training set and test sets SALIGN and Prosop are 18% and 20%, respectively.

Table 1 shows the alignment accuracy of different methods given by different benchmarks along with the standard errors estimated by bootstrap simulation on 10,000 re-sampling of the data. There is a consistent improvement from SP$^3$, SP$^4$ to SP$^5$. The absolute changes range from 1.9% to 2.4% (3.4%) from SP$^3$ (SP$^4$) to SP$^5$ while the relative increases are between 3–5% (5–6%) [SP$^3$ relative to SP$^4$ (SP$^5$)]. These changes are significantly greater than the estimated standard errors. The improvement is remarkable considering the fact that ProSup benchmark was used as the training set to optimize the parameters of SP$^3$ [11] and SP$^4$ [12].
the comparison within SP methods, we used original sequence profiles from Ref. [11].

Table 2 indicates that the improvement over SP$^3$ and SP$^4$ in success rate of fold recognition by SP$^5$ exists in all three levels (family, superfamily, and fold). The largest improvement over SP$^4$ is observed in fold level (7% absolute increase in Top 1 and 5% absolute increase for the best in Top 5; 22% relative increase in Top 1, 9.5% in top 5). This is somewhat expected because the method was trained for remote homolog recognition (structurally similar protein with less than 30% sequence identity, PREFAB benchmark). Again the relative improvement of SP$^5$ over SP$^3$ and SP$^4$ is significantly larger than the standard errors estimated from bootstrap simulations. We further removed 43 proteins that have >30% sequence identity with the training sequences in the PREFAB benchmark. Their effect on the final result is negligible. For comparison, we also include the results of PSI-BLAST [9], SPARKS [10], HHsearch / HHpred [27] and FOLDpro [48]. The performance of SP$^3$, SP$^4$ and FOLDpro was from Ref. [10] and Ref. [48], respectively. We further performed PSIBLAST and HHpred locally with their default parameters.

Above success rates of matching sequences within the same SCOP classification are based on somewhat subjective SCOP definition of family, superfamily and fold [49]. A more direct measurement of accuracy is to calculate the accuracy of the first-definition of family, superfamily and fold [49]. A more direct measurement of accuracy is to calculate the accuracy of the first.

### Table 2. The success rate for recognizing proteins within the same family, superfamily, or fold in the Lindahl benchmark.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Family only (%)</th>
<th>Superfamily only (%)</th>
<th>Fold only (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 1</td>
<td>Top 5</td>
<td>Top 1</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>62.4</td>
<td>67.6</td>
<td>16.0</td>
</tr>
<tr>
<td>SPARKS$^5$</td>
<td>81.6</td>
<td>88.1</td>
<td>52.5</td>
</tr>
<tr>
<td>HHpred</td>
<td>82.9</td>
<td>87.1</td>
<td>58.8</td>
</tr>
<tr>
<td>FOLDpro$^5$</td>
<td>85.0</td>
<td>89.9</td>
<td>55.5</td>
</tr>
<tr>
<td>SP$^{1,4}$</td>
<td>81.6 ± 0.07</td>
<td>86.8 ± 0.06</td>
<td>55.3 ± 0.11</td>
</tr>
<tr>
<td>SP$^{4,6}$</td>
<td>80.9 ± 0.07</td>
<td>86.3 ± 0.06</td>
<td>57.8 ± 0.11</td>
</tr>
<tr>
<td>SP$^{5,6}$</td>
<td>82.4 ± 0.07</td>
<td>87.6 ± 0.06</td>
<td>59.8 ± 0.11</td>
</tr>
</tbody>
</table>

*The percentage in each cell is the fraction of correctly recognized match of proteins in the same fold, superfamily, and family as first rank or within top 5 rank of the template.

1. From Ref. [10].
3. This work.
4. This work (The 43 proteins with >30% sequence similarity to PREFAB training set are removed).
Recent progress in sequence alignment and structure prediction has suggested the importance of variable gap penalties in protein sequence alignment [51]. Different form of context (either structure or sequence context or both)-dependent gap-penalty model has been proposed [52,53]. Employing fractional-powered gap insertion/deletion profiles is another new feature introduced in SP5. While these insertion/deletion profiles were used, previously [26–28], our trial-and-error analysis indicates that the fractional-powered gap insertion/deletion profiles with a power of 0.1 seem to be more suitable for improving alignment accuracy. However, more systematic comparative studies are needed to check if any other functional forms are more appropriate.

To analyze the usefulness of the new gap model, we made a version of SP5 with the previously used gap model and found that new gap model leads to a small but positive increase in alignment accuracy (0.5% in PREFAB, 1.5% in ProSup and 0.1% in SALIGN). Thus, the main contribution for improved ability in fold recognition by SP5 is due to introduction of torsion angles.

SP3 and SP4 were among the top performers in automatic servers in CASP 6 and 7 [13,12]. It is noted that in CASP7, SP3 scored higher than SP4 according to GDT-HA, TMscore, and AL0 for all targets. A close examination [12] indicates that SP3 is slightly more accurate than SP5 in hard targets (FM category), but slightly worse than SP3 in other targets (TBM category). This is perhaps because all parameters were optimized for fold recognition targets. On the other hand, SP5 performs consistently better than SP3 at both FM and TBM categories if the cumulative Z-score is used [12]. The development of SP5 continues our emphasis on searching a more sensitive method for fold recognition. Significant improvement of SP5 over SP4 and SP3 indicates that SP5 is among the most accurate automatic servers for fold recognition.

In the SP serial methods, the alignment generated for fold recognition is used directly in modeling. It is quite possible that a separate alignment method optimized for modeling may further improve the accuracy of predicted model. This will be a subject of future studies.

**Methods**

**Alignment Score**

The alignment score of SP5 for aligning query position i with the template position j is

\[
S(i,j) = -(1 - w_{\text{struc}}) F_{\text{seq}}(i,j) \cdot M_{\text{template}}(j)
- w_{\text{struc}} F_{\text{template}}(j) \cdot M_{\text{seq}}(i,j)
- w_{\text{2ndary}} \delta(i,j) - w_{\text{sa}} (1 - 2 |sa(i) - sa(j)|)
- w_d (1 - d/90) + s_{\text{shift}}
\]

with four weight parameters \(w_{\text{struc}}, w_{\text{2ndary}}, w_{\text{sa}}, \text{ and } w_d\) and a constant shift \(s_{\text{shift}}\). This score represents weighted matching of five profiles that are described in detail below.

The first term in Eq. (1) is the profile-profile comparison between the sequence profile from the query sequence and that from the template sequence. \(F_{\text{seq}}(i,j)\) is the sequence-derived frequency profile of the query sequence. \(M_{\text{template}}(j)\) and \(M_{\text{seq}}(i,j)\) are the sequence-derived log odd profile of the template sequence and that of query sequence, respectively. These sequence profiles are constructed by three iterations of PSIBLAST [9] searching (E value cutoff 0.001) against non-redundant (NR) sequence database, which was filtered to remove low-complexity regions, transmembrane regions, and coiled-coil segments [29].
The second term in Eq. (1) compares the sequence profile from the query sequence and that derived from the template sequence (sequence profiles that would “fit” to the structure). $F_{template}(j)$ is a depth-dependent sequence profile generated from the sequences of those structural fragments that are similar to 9-residue segment structures of the template [11].

The third term in Eq. (1) measures the difference between the predicted secondary structure of the query sequence and the actual secondary structure of the template. $\delta_{sa}(i)$ is a simple function of the secondary structure element $\alpha$ of the query at sequence position $i$ and $\beta$ of the template at sequence position $j$. $\delta_{sa}(i) = 1$ if $\alpha = \beta$ and $\delta_{sa}(i) = -1$ if $\alpha \neq \beta$. We use a three-state definition of secondary structures (H for helix, E for strand, and C for coil). The secondary structure of query sequences is predicted by SPINE [31]. The first three terms constitute the method SP5 [11] except that PSI-PRED [29] rather than SPINE [31] was used in SP5 to predict the secondary structure of the query sequence. DSSP [30] is used for analyzing template structures because SPINE was trained based on the DSSP definition of secondary structures.

The fourth term in Eq. (1) is the matching score between the predicted solvent accessibility of the query sequence and solvent accessibility of the template structure. $sa(i)$ and $sa(j)$ are the predicted residue solvent accessibility of query sequence and that of the template structure, respectively. The residue solvent accessibilities of query sequence are predicted by Real-SPINE [25] while residue solvent accessibility of template structures are calculated from DSSP [30] and normalized by unfolded solvent accessible surface areas [32]. The fourth term meets the method SP3 [33] except that in SP3, PSI-PRED [29] rather than SPINE [31] was employed to predict the secondary structure of the query sequence, and the real values of solvent accessibility from Real-SPINE [25] rather than two-state classifications by SABLE [34] are used to predict the residue solvent accessibility of the query sequence.

The fifth term in Eq. (1) is a new addition in SP5. It characterizes the difference between predicted angles ($\psi(i)$ and $\phi(i)$) of the query sequence and actual angles ($\psi(j)$ and $\phi(j)$) of the template structure with

$$\Delta = \frac{1}{2} \left[ (\psi(i) - \psi(j))^2 + (\phi(i) - \phi(j))^2 \right]$$

Real values of angles for the query sequence are from Real-SPINE 2.0 [24] while these angles are calculated by DSSP [30] for the template structure. Real-SPINE 2.0 is a method for real-value prediction of torsion angles by using back-propagation neural networks trained with a sliding 21-residue window of sequence profiles, representative amino acid properties, and predicted secondary structures. The ten-fold-cross-validated mean absolute errors are $38^\circ$ for $\psi$ and $25^\circ$ for $\phi$, respectively.

Profile-based Gap Model

SP5 [11] and SP4 [12] employ a simple secondary-structure dependent gap penalty. No gaps are allowed if $s_i = s_j = \alpha$ (helix) or $s_i = s_j = \beta$ (sheet). The gap opening ($a(i)$) and gap extension ($a(k)$) penalties are applied to other regions. In this paper, we construct a profile-based gap model from the multiple sequence alignment made by PSIBLAST [9]. The multiple sequence alignment allows us to calculate the probability of deletion at sequence position $i$, $P_{del}(i)$, and the probability of insertion at sequence position $i$, $P_{ins}(i)$, $P_{del}(i) = n_{del}^i / N$ and $P_{ins}(i) = n_{ins}^i / N$ where $n_{del}^i$, $n_{ins}^i$, and $N$ are number of deletions in sequence position $i$, number of insertions in sequence position $i$, and total number of aligned sequences, respectively.

Thus, we have four profiles: two for query sequences and two for template sequences ($P_{del}^{query}(i)$, $P_{ins}^{query}(i)$, $P_{del}^{template}(i)$, and $P_{ins}^{template}(i)$). The gap penalty is calculated as follows. We still use $\omega_{ga}$ as the gap opening penalty. The extension gap penalty is modified by $1 - \left( \frac{P_{del}^{query}(i)}{P_{del}^{template}(i)} \right)^2 + \left( \frac{P_{ins}^{query}(i)}{P_{ins}^{template}(i)} \right)^2$ for residue $i$ in the query sequence that is aligned with a gap after residue $j$ in template. Similarly, the extension gap penalty is modified by $1 - \left( \frac{P_{del}^{query}(i)}{P_{del}^{template}(i)} \right)^2 + \left( \frac{P_{ins}^{template}(i)}{P_{ins}^{query}(i)} \right)^2$ or residue $j$ in template that is aligned with a gap after residue $i$ in query. Here, $\omega_{ga}$ is a to-be-optimized weight factor. Usually, $lnP_{del}^{query}/template(j)$ is an energetic term. Here, we use $\omega_{g} = P_{del}^{query}/template(j)$ rather than $lnP_{del}^{query}/template(j)$ to avoid singularity at $P_{del}^{query}/template(j) = 0$. We set $\gamma = 0.1$ by trials and errors.

Dynamic Programming and Template Ranking

Similar to SP3 and SP5, we used the Smith-Waterman local alignment algorithm [35] to optimize the score that matches the query profiles with template profiles based on Eq. (1) with the revised gapping method described above. Note that the optimization of alignment is to minimize the total alignment score due to the negative signs in Eq. (1).

The templates are ranked based on the difference score between the raw alignment score and the reverse alignment raw score in which the alignment is made with the reversed query sequence [36]. The results of fold-recognition alignment are used to build Cα models based on native template structure. This is done by directly transferring the Cα coordinates of the template structures to the aligned residues in the query sequence. If there is no structural similarity between first two models (defined as zero MaxSub score [37]), templates will be re-ranked by the greater one of 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 $Z(i) = \left( \frac{S_{\alpha}(i) - S_{\alpha}^c}{\sigma} \right)$, where $S_{\alpha}$ and $\sigma$ denotes the average and standard deviation of normalized score for all the templates. This ranking mechanism was based on an empirical observation. We found that ranking based on the difference score between the raw alignment score and the reverse alignment raw score works well only if there is some structural similarity between the top-two ranked models (i.e. a significant structural cluster detected). Otherwise, ranking based on Z-scores works better [11].

Author Contributions

Conceived and designed the experiments: YZ. Performed the experiments: WZ SL. Analyzed the data: WZ SL. Wrote the paper: YZ WZ SL.

References


