Template Scoring Methods for Protein Torsion Angle Prediction

Zafer Aydin[1(✉)], David Baker[2], and William Stafford Noble[3]

[1] Department of Computer Engineering, Abdullah Gul University, 38080 Kayseri, Turkey
zafer.aydin@agu.edu.tr
[2] Department of Biochemistry, University of Washington, Seattle, WA 98195, USA
[3] Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA

Abstract. Prediction of backbone torsion angles provides important constraints about the 3D structure of a protein and is receiving a growing interest in the structure prediction community. In this paper, we introduce a three-stage machine learning classifier to predict the 7-state torsion angles of a protein. The first two stages employ dynamic Bayesian and neural networks to produce an ab-initio prediction of torsion angle states starting from sequence profiles. The third stage is a committee classifier, which combines the ab-initio prediction with a structural frequency profile derived from templates obtained by HHsearch. We develop several structural profile models and obtain significant improvements over the Laplacian scoring technique through: (1) scaling templates by integer powers of sequence identity score, (2) incorporating other alignment scores as multiplicative factors (3) adjusting or optimizing parameters of the profile models with respect to the similarity interval of the target. We also demonstrate that the torsion angle prediction accuracy improves at all levels of target-template similarity even when templates are distant from the target. The improvement is at significantly higher rates as template structures gradually get closer to target.

1 Introduction

Protein 3D structure prediction benefits greatly from prediction of various 1D and 2D structural attributes such as secondary structure, backbone torsion (dihedral) angles, solvent accessibility, disordered regions, and contact maps [6]. Methods that predict structural properties of proteins typically employ sequence-based frequency profiles in their feature sets to utilize information in similar proteins. These profiles can be in the form of position specific scoring matrices (PSSM) or hidden Markov models (HMM) and can be derived by aligning the amino acid sequence of the query with sequences in a large protein database using an efficient algorithm such as PSI-BLAST [1] or HHBlits [16]. Despite the many efforts for improving the quality of sequence-based alignments and their profiles, the accuracy of 1D and 2D predictions has come to saturation due to the difficulty of eliminating false positives especially when the query sequence diverges from those in the protein database considerably. Recently, there has been a growing interest in using

A. Fred et al. (Eds): BIOSTEC 2015, CCIS 574, pp. 206–223, 2015.
DOI: 10.1007/978-3-319-27707-3_13

structural profiles as input features for predicting various structural characteristics of proteins. A structural frequency profile is a position specific scoring matrix (PSSM) that is constructed from the structural labels of templates (*i.e.*, hit proteins) obtained by aligning the target (*i.e.*, query) against a set of proteins. To date structural profiles have been derived mainly for protein secondary structure [7,12]; backbone structural motifs, solvent accessibility, contact density [13]; and shape strings [23].

Among the various structural properties of proteins, predicted secondary structure has been widely employed by many structure prediction methods to reduce the conformational space that must be explored. One limitation of using secondary structure is the inability to impose constraints for loop (or coil) segments, which do not have a well-defined structure [30]. On the other hand, predictions of backbone torsion angles can provide powerful and more fine-grained restraints as compared to secondary structure. It is anticipated that accurately predicted torsion angles will one day replace the dominant role played by secondary structure in tertiary structure prediction [15].

Despite the variety of methods proposed for predicting backbone torsion angles of proteins [4,8,17,18,22,27], less effort has been made to systematically incorporate structurally related templates into torsion angle predictions [13]. Furthermore, to the best of our knowledge, there is no work in the literature that inspects the accuracy of torsion angle predictions at all levels of target-template similarity (*i.e.*, from easy to difficult targets).

One approach for incorporating template information is to construct a frequency profile matrix, in which the occurrence frequencies of template residues are accumulated followed by a normalization step. This matrix is later included to the feature set of a machine learning classifier, which predicts the structural class labels of the amino acids. Methods that have been developed for profile matrices mainly use Laplacian counts, which is a technique that gives equal weights to templates [12]. As an alternative to the Laplacian count method, a new scoring technique has been proposed which scale the templates by the third power of the sequence identity score and the structural quality factor [14,24].

In this paper, we propose new structural profile methods that incorporate various score terms of HHsearch alignments [19] to predict torsion angles of proteins. We combine structural profiles with ab-initio predictions obtained from a two-stage classifier using a linear committee approach. Our method is able to generate specific and effective predictions for targets at all difficulty levels. We achieve this by optimizing certain parameters of profile models with respect to the similarity interval of the target.

2 Methods

2.1 Backbone Torsion Angles

Each amino acid residue has three associated backbone torsion angles: ϕ, ψ, and ω. The angle ϕ denotes rotation about the C_α-N bond of the residue, ψ denotes rotation about the bond linking C_α and the carbonyl carbon, and ω

denotes rotation about the bond between the carbonyl carbon of the current residue and the nitrogen of the next residue. We compute ϕ, ψ, and ω from the 3-D coordinate information in Protein Data Bank (PDB), which is the database of solved protein structures. Each of these angles is constrained to the range $[-180, 180]$.

Following [5], we first subdivided residues into five torsion angle classes, which represent the major clusters observed in PDB. However, to reduce the imbalance in the sizes of these classes, we further subdivided the two most common labels (A and B) according to whether the secondary structure class is loop or not. The resulting seven labels are described in Table 1.

Table 1. 7-state torsion angle labels. The table lists the seven torsion angle classes, their definitions, and the percentage of residues assigned to each class in the PDB-PC90 data set. ss denotes the secondary structure label of the amino acid residue.

Label	Definition	Percent
L	$\lvert\omega\rvert \geq 90$, $\phi < 0$, $-125 < \psi \leq 50$, ss = loop	11.94
A	$\lvert\omega\rvert \geq 90$, $\phi < 0$, $-125 < \psi \leq 50$, ss $\neq$ loop	38.21
M	$\lvert\omega\rvert \geq 90$, $\phi < 0$, $\psi \leq -125$ OR $\psi > 50$, ss = loop	20.08
B	$\lvert\omega\rvert \geq 90$, $\phi < 0$, $\psi \leq -125$ OR $\psi > 50$, ss $\neq$ loop	22.27
E	$\lvert\omega\rvert \geq 90$, $\phi \geq 0$, $\lvert\psi\rvert > 100$	1.92
G	$\lvert\omega\rvert \geq 90$, $\phi \geq 0$, $\lvert\psi\rvert \leq 100$	4.73
O	$\lvert\omega\rvert < 90$	0.84

2.2 Torsion Angle Class Prediction

Based on the definition given in Table 1, the 7-state torsion angle prediction problem can be stated as follows. For a given protein, the goal is to assign to each amino acid a torsion angle label from the alphabet $\{L, A, M, B, E, G, O\}$ as shown in Fig. 1.

```
LWGLVKQGLKCEDCGMNVHHKCREKVANLC
MMELMGLBBBBLLLGMBBMAAAALLMMLMO
```

Fig. 1. 7-state torsion angle class prediction problem. The first row shows the amino acid sequence of the target and the second row is the sequence of 7-state torsion angle labels, which are defined according to Table 1.

2.3 Alignment Methods

Deriving Templates for Structural Profiles. In this paper, we used the HHsearch method [19] to detect the templates that are similar to a given target.

HHsearch first derives an HMM-profile for the target and aligns it against a database of HMM-profiles [19]. At the end of the alignment, it ranks the templates (*i.e.* hits) according to a probability score ranging from 0 % to 100 % and reports the ones that score above a threshold. An example alignment is shown in Fig. 2. We used the following commandline to compute HMM-HMM alignments for each target: `./hhsearch -i protein.hhm -d hhm3 -o protein.hhr -cpu 2 -mact 0.05 -ssw 0.11 -atab protein. start.tab -realign -E 100 -cov 20 -b 20`. We then selected the HMM-HMM alignments that score above the given threshold as the templates. Note that, HHsearch uses predicted secondary structure to be able to compute sensitive HMM-HMM alignments. We used the PSIPRED version 2.61 [11] to predict secondary structures. All these alignments were generated in 2011. Further details on HHsearch and the HMM-HMM alignments can be found in the corresponding documentation [20,21].

```
No 3
>3fy3A
Probab=100.00  E-value=3.6e-41  Score=236.01  Aligned_columns=201  Identities=22%

Q ss_pred            CCEEECCC---EEE--BCCCCCEEEEECCCC---EEEE-CHHCCCCCCCEEEEC-------------------CCCHHHHH
Q ss_conf            99896176---479--87898379944863---4985-133664889879976-------------------88256532
Q 2odlA.fasta      4 GMDVVHGT---ATM--QVDGNKTIIRNSVD---AIIN-WKQFNIDQNEMVQFL-------------------QENNNSAV   55
(372)
Q Consensus        4 g~~v~~g~~~~~i~~~~~~~~i~q~s~~~~~~~n~w~sFnIg~~~~v~f-------------------q~~~a~vi   55
(372)
                      .+.|+.+.   -.+  ..++.+.|+..+|.   ..|| |++|||++.+.+.+|                   ...+|++|
T Consensus        1 ~gIv~~~~~~~~v~~~nG~~vinI~~Pn~~GiS~N~y~~FnV~~~G~vlNNs~~~~~t~l~G~i~~NpnL~~~~A~~I   80
(234)
T 3fy3A            1 NGIVPDAGHQGPDVSAVNGGTQVINIVTPNNEGISHNQYQDFNVGKPGAVFNNALEAGQSQLAGHLNANSNLNGQAASLI   80
(234)
T ss_dssp            CCCEECSSTTCCEEEEETTTEEEEECCCCCTTSEEEEEEEECCBCTTCEEEEECSSCEEETTTEEECCCTTSSSCCCSEE
T ss_pred            CCEEECCCCCCCCEEEECCCCCEEEEEECCCCCCCCEEEEEEEEECCCCCEEEECCCCCCCCCCCCCCEECCCCCCCCCCEE
T ss_conf            95663798988677755999328986678888534532012103997599657554530002022157533787642 89
```

Fig. 2. An example HMM-HMM alignment obtained by HHsearch.

Generating Position-Specific Scoring Matrices for The Ab-Initio Method. We employed PSSMs generated by the PSI-BLAST [1] and HHMAKE [19] algorithms as input features. We used BLAST version 2.2.20 and the NCBI's non-redundant (NR) database dated June 2011 to generate PSI-BLAST PSSMs. We generated the HMM-profiles by HHsearch version 1.5.1 [19]. Note that in deriving the HHMAKE PSSMs we did not perform any HMM-HMM alignments. After deriving PSSMs we scaled them to the interval $[0, 1]$ by applying a sigmoidal transformation. Detailed descriptions of the PSSMs and the sigmoidal transformation can be found in [2].

2.4 Structural Frequency Profiles for 7-State Torsion Angles

Our structural profile is a $7 \times N$ matrix where rows represent the torsion angle classes and columns denote the amino acids of the target. An example structural profile is illustrated in Fig. 3.

The entries of a structural frequency profile matrix represent the propensity of having a particular torsion angle class at a given amino acid residue of the target protein. It therefore acts as a signature summarizing the 7-state torsion label expectancy of the target residues. The entries of this matrix are normalized

	1	2	...	N
L	0.40	0.20	...	0.18
A	0.17	0.30	...	0.02
M	0.01	0.13	...	0.08
B	0.03	0.07	...	0.12
E	0.19	0.11	...	0.04
G	0.16	0.09	...	0.06
O	0.04	0.10	...	0.50

Fig. 3. A structural frequency profile for 7-state torsion angle representation. Rows represent the torsion angle classes and columns denote the amino acids of the target. Each column sums to 1.

to the interval [0,1] and each column sums to 1 similar to a marginal *a posteriori* probability distribution. For this reason, we denote a structural profile by $P_s(t_j|y)$, where t_j is the torsion angle class of the j^{th} residue and y is the amino acid sequence of the target.

A structural profile can be obtained by collecting the occurrence frequencies of the structure labels of the template proteins. There could be various approaches for computing a structural frequency profile. The following sections explain the methods that have been implemented in this paper.

Laplacian Counts. The most straightforward approach for deriving a structural frequency profile is to count the occurrence frequencies of torsion angle labels that match to a given position of the target, which is followed by a normalization step. The Laplacian count method is employed by most of the approaches that have been proposed for computing structural profiles to date [12]. First a count matrix C is obtained that contains the occurrence frequencies of the structure labels of the templates, which is formulated as follows

$$C(i,j) = \sum_{A(j,k)} \delta(T(j,k),i) \tag{1}$$

where $C(i,j)$ is the $(i,j)^{th}$ entry of the count matrix such that $i \in \{L,A,M,B,E,G,O\}$ is the torsion angle class, j is the residue position of the target, $A(j,k)$ is the residue of the k^{th} database protein aligned to the j^{th} position, $T(j,k)$ is the corresponding torsion angle class of the template, and $\delta(T(j,k),i)$ is the Kronecker delta function defined as

$$\delta(t,i) = \begin{cases} 1 \text{ if } t = i \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

In other words, each torsion angle label that is aligned to the j^{th} position of the target contributes by a count of 1, which is also known as the Laplacian count method. Once the count matrix is obtained it is normalized so that each column sums to 1. This is formulated as

$$M_a(i,j) = \begin{cases} \frac{C(i,j)}{\sum_i C(i,j)} & \text{if } |A(j)| > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $A(j)$ is the set of all residues aligned to the j^{th} residue of the target, $|A(j)|$ is the number of residues in $A(j)$, and M_a is the normalized count matrix. In this formulation, the residues of the target are divided into two categories. In the first group, we have "aligned" positions (represented by the condition $|A(j)| > 0$) where at least one residue is aligned from a database protein and in the second group there is the set of "unaligned" positions (*i.e.*, the case where $|A(j)| = 0$) for which no residues are aligned from any hits. The second condition is realized for positions that correspond to gapped regions and for positions that are left out of the aligned regions when a local alignment algorithm is employed.

After the normalization step, the structural profile matrix can be computed as

$$M(i,j) = \begin{cases} M_a(i,j) \text{ if } |A(j)| > 0 \\ M_b(i,j) \text{ otherwise} \end{cases} \tag{4}$$

where $M_b(i,j)$ is the background probability of aligning a template residue with torsion class i to the j^{th} residue of the target. In this paper, we use predictions from the ab-initio classifier for the background distribution of torsion angle labels.

Weighing Hits by Integer Powers of Sequence Identity Scores. A second method for computing structural profiles weights templates by integer powers of the sequence identity score. In HHsearch, this score is computed for each target template pair and is represented by the "Identities" field as shown in Fig. 2. We first divide this score by 100 and convert it to a weight value. We then compute an integer power of this weight, which is used to scale templates that contribute to the structural profile. This is expressed in the equations below

$$C(i,j) = \sum_{A(j,k)} \theta(T(j,k), i) \tag{5}$$

$$\theta(t,i) = \begin{cases} I^a \text{ if } t = i \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

where C is the count matrix, θ is the new occurrence count function replacing the Kronecker delta in (2), I is the sequence identity score of the k^{th} template and a is an integer that represents the strength of the amplification one wishes to impose on the structurally similar templates. The remaining terms are the same as their counterparts in (1) and (2). This type of template scaling has two benefits. The first one is related to scaling templates by sequence identity scores, which increases the contribution of structurally closer templates while reducing the votes of distant ones. The second benefit comes by taking integer powers of I, which manages the situation where a handful of structurally similar templates are followed by many less similar or distant templates. In such a scenario, if we

use the Laplacian counts as in Sect. 2.4 or weigh templates by sequence identity scores only (*i.e.*, $a = 1$) the contribution of the similar templates would be suppressed by many structurally less similar candidates. To further amplify the effect of structurally similar templates and to reduce the contribution of false positives (*i.e.*, noise) it is useful to take integer powers of the sequence identity scores as formulated in (6). Once we compute the count matrix, we normalize it as in (3). All the other steps in deriving the structural profile are the same as in Sect. 2.4.

Incorporating Quality of Templates. In addition to taking integer powers of the sequence identity score, it is also possible to include other weight factors to the score function in (6). One such measure assesses the experimental quality of the templates and is proposed in [14,24]. When we employ this approach to score the templates, (6) takes the following form:

$$\theta(t,i) = \begin{cases} \frac{I^a}{q} & \text{if } t = i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where q is the quality of the template computed as X-ray resolution + R-factor / 20 as proposed in [10]. According to this measure, a template with a higher experimental quality has a lower q parameter. Since this measure requires the X-ray resolution of the template, we apply it to those templates that have been solved by the X-ray method only ignoring the remaining templates for the target.

Incorporating Other Alignment Scores. When two proteins are aligned to each other, typically several score terms are calculated for assessing the statistical significance including e-value, raw similarity score, and percentage of sequence identity. Employing these terms in scaling the templates could also be useful in constructing a structural profile. With this motivation, we first incorporated the e-value score into the occurrence count function by converting it to a multiplicative weight factor as in [28]. This is formulated as

$$\theta(t,i) = \begin{cases} w_e\, I^a & \text{if } t = i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where w_e is the e-value weight defined as

$$w_e = \begin{cases} 1 & \text{if } E < 10^{-10} \\ -0.05 \log_{10}(E) + 0.5 & \text{if } 10^{-10} \leq E < 10^{10} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

such that E is the E-value of the alignment. Note that we dropped the quality term as it did not bring any significant benefits, which is verified by our simulations. According to (9), w_e is set to 1 when the target-template similarity is above a certain threshold ($E - value < 10^{-10}$) and decreases linearly as the E-value of the target-template alignment is greater than 10^{-10} until it becomes considerably high (*i.e.*, 10^{10}), in which case it is set to zero.

In addition to the E-value, we also considered incorporating the overall raw similarity score of the alignment into our structural profiles. For this purpose, we modified the occurrence count function as

$$\theta(t,i) = \begin{cases} s\, w_e\, I^a & \text{if } t = i \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where s is the raw score of the alignment. For HHsearch, this is the overall similarity score obtained at the end of the HMM-HMM alignment, which is computed as the sum of the similarities of the aligned profile columns minus the gap penalties [19]. A slight variation of this approach normalizes the raw score with the length of the aligned region as

$$\theta(t,i) = \begin{cases} \frac{s}{L}\, w_e\, I^a & \text{if } t = i \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

such that L is the length of the aligned region and is given as the field denoted as "Aligned_columns" in HHsearch's output (see Fig. 2).

Scaling Columns of the Alignment. Up to this point, we scaled the templates uniformly throughout the aligned positions without discriminating the individual columns of the alignment. In this section, we explain an approach for amplifying local regions within an alignment that could potentially contribute more accurate torsion label information and suppressing those that could be locally more distant. For this purpose, we include the similarity score between the aligned residues from a BLOSUM matrix into the occurrence count function as formulated below

$$\theta(t,i) = \begin{cases} \frac{s}{L}\, w_e\, I^a\, e^b & \text{if } t = i \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where b is the similarity score such as BLOSUM matrix score between the j^{th} residue of the target and the residue in the template that is aligned to the target residue. When the two residues are biologically similar to each other we expect this score term to be larger than the term obtained from dissimilar pairings. This approach has the potential to amplify local matches between motifs that are common both in the target and the template. Note that normalizing the sequence alignment score with L is optional. For instance, a slightly modified version of (12) does not perform such type of normalization:

$$\theta(t,i) = \begin{cases} s\, w_e\, I^a\, e^b & \text{if } t = i \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Selecting Templates Within a Score Window. To improve the speed of predictions, it is possible to select those alignments that score above a given

HHsearch probability threshold denoted as s_T. Let s_B represent the probability score of the top scoring alignment. Then, s_T is computed as

$$s_T = s_B - W \tag{14}$$

where W is the probability score window, which is a hyper-parameter of the template based predictor.

2.5 Prediction Model

Ab-Initio Predictor. Our *ab-initio* torsion class predictor is a hybrid architecture, in which four dynamic Bayesian network (DBN) models are combined with a neural network. Two of the DBNs use PSSMs derived by PSI-BLAST [1] and the other two use PSSMs from the HHMAKE module of the HHsearch method [19]. Details of the ab-initio predictor can be found in [2,3]. For simplicity we treat the output signal of the neural network as a probability distribution due to the constraints it satisfies (*i.e.*, it sums to 1 and takes values from 0 to 1). This distribution is denoted as $P_a(t_j|x)$, which represents the *ab-initio* likelihood of the j^{th} residue to have t_j as the torsion angle label given x, the set of input features around position j. Hence our neural network predicts the torsion angle label of the residue at the center of the feature window by selecting the particular label with the maximum discriminant score at the output layer. This is formulated as

$$t_j^* = \arg\max_{t_j} P_a(t_j|x). \tag{15}$$

Committee Predictor. The committee predictor combines the ab-initio predictions of torsion angle classes with the structural frequency profile according to the following equation:

$$P_c(t_j|x,y) = \begin{cases} (1-\lambda)P_a(t_j|x) + \lambda P_s(t_j|y) & \text{if aligned} \\ P_a(t_j|x) & \text{if unaligned} \end{cases} \tag{16}$$

where $P_c(t_j|x,y)$ is the combined likelihood of having torsion angle label t_j for the residue at position j, λ is the weight assigned to the structural profile, x is the feature set of the ab-initio predictor described in Sect. 2.5, y is the amino acid sequence of the target, $P_a(t_j|x)$ is the distribution of torsion angle classes obtained from the ab-initio predictor, and $P_s(t_j|y)$ is the structural profile computed from the templates. According to this equation, the combined likelihood is the weighted average of the ab-initio predictions and the structural profile for positions that are aligned to at least one template residue and it becomes equal to the likelihood of the ab-initio predictor only for the remaining positions.

After computing $P_c(t_j|x,y)$, we predict the torsion angle class of the j^{th} residue as the particular label that maximizes $P_c(t_j|x,y)$ as

$$t_j^* = \arg\max_{t_j} P_c(t_j|x,y), \tag{17}$$

where t_j^* is the final prediction of the committee method.

2.6 Optimizing Model Parameters

We optimized the weight λ in (16) that is used to combine the structural profiles with the ab-initio predictor, the probability score window W in (14) that is used to select the HHsearch templates within a window and the power parameter a in (6). For this purpose, we followed an iterative strategy in which we select the optimums that yield the best overall accuracy. In the first iteration, we optimized λ, followed by W and a. In the second and third iterations, we refined our estimates by repeating the optimization of λ, W and a in sequence. At each optimization step, we set the remaining two parameters to their optimums obtained up to that point. We performed this optimization on the validation set described in Sect. 2.7 using the structural profile model in (6) together with the windowing approach in (14). Once we optimize these parameters, we trained the ab-initio predictor model on the set of 4205 proteins (i.e. the training set in Sect. 2.7) and computed predictions using the optimums.

2.7 Datasets

PDB-PC90 Dataset. To obtain the PDB-PC90 dataset, we used the PISCES server [25,26] with the following set of criteria: percent identity threshold of 90 %, resolution cutoff of 2.5 Å, and R-value cutoff of 1.0. We also used PISCES to filter out non-X-ray and C_α-only structures and to remove short (< 30 amino acids) and long (> 10000 amino acids) chains. This dataset contained 17056 chains.

Training, Validation and Test Sets. We randomly selected 5161 proteins from the PDB-PC90 dataset. Among those, we randomly selected a set of 994 proteins for the first test set. From the set of 5161 proteins, we then removed those proteins that are similar to the test set using a 10 % sequence identity threshold. The remaining set contained 4205 chains, which is used to train our ab-initio prediction method. We computed the HHsearch alignments for the set of 994 proteins, which are used for computing the structural profiles of torsion angle classes and for predicting the torsion angle classes.

We further split the test set into two each containing 497 proteins. The first half is used as a validation set to optimize selected parameters of the model and the second is used to evaluate the prediction accuracy when the optimum parameters are employed during profile construction.

Similarity Intervals and Subsets of the Test Set. To distinguish easy targets from difficult ones, we defined similarity intervals using the HHsearch alignments from half of the proteins in our test set (see Sect. 2.7). For each target, we first selected the maximum sequence identity score from the set of target-template alignments. Then we ranked those scores and defined percentile intervals of sequence identity with increments of 5 %. This initially produced a total of 20 intervals. We combined the eighth and ninth intervals as the maximum sequence identity scores for those targets were very close to each other. We also combined the tenth up to the twentieth intervals since the maximum sequence identity score was 100 % for all the targets in those bins. This procedure resulted in a total of 9 sequence identity intervals. In the last step, we further reduced the number of intervals to 5 by combining the 2nd and 3rd, 4th and 5th, and 7th up to 9th intervals. The resulting intervals are tabulated below

Table 2. Intervals of sequence identity scores. The intervals are defined by selecting the target-template alignments with maximum sequence identity scores followed by sorting these scores in ascending order. Percentile increments of 5 % results in a total of 20 bins, which are further reduced to 5 intervals.

Interval	Percentiles (%)	Max Identity (%)
Low	0–5	0.0–26.0
Medium-Low	5–15	26.0–35.0
Medium	15–25	35.0–50.0
Medium-High	25–30	50.0–80.0
High	30–100	80.0–100.0

According to Table 2, the first interval represents targets with the most distant templates (*i.e.*, those with the maximum sequence identity score of 26 % or less) and the last interval represents targets that contain highly similar templates (*i.e.*, those with the maximum sequence identity score greater than 80 %). Based on this binning, we further divided our test set of 994 proteins (see the previous section) into 5 subsets such that each contained those targets that fall into one and only one of the intervals defined in Table 2. The number of proteins and amino acids in each of these subsets are summarized in Table 3.

Note that the number of proteins in each subset is not uniformly the same (especially true for the last set that contains targets from the "High" category) mainly because datasets have been constructed by random sampling from PDB without enforcing specific constraints for having equal number of samples in each interval. Nonetheless we have enough samples in each subset mainly because the torsion angle class prediction is performed on each residue separately. Furthermore the proportion of target positions that are aligned to at least one template residue is considerably high. This shows that a structural profile column is computed using the aligned templates for most of the target residues.

Table 3. The five subsets of the test set with 994 proteins. The number of proteins, the total number of amino acids and the number of amino acids that are aligned to at least one template residue are shown for each subset. The subsets are derived based on the intervals defined in Table 2.

Subset	# proteins	# residues	# aligned res
Low	56	12903	12665
Medium-Low	99	21792	21682
Medium	95	22596	22561
Medium-High	62	15326	15295
High	682	160037	159993
Total	994	232654	232196

We did a similar partitioning for the validation and the second test set where each subset contained proteins from the corresponding similarity interval. As a result of this procedure the validation subsets contained 25, 51, 52, 24, and 345 proteins and the second test set contained 31, 48, 43, 38, and 337 proteins respectively starting with the lowest similarity up to the highest. To reduce the computation time of optimizations in the high similarity interval we only used 35 proteins instead of 345.

3 Results

3.1 Accuracy of Structural Profiles only

We first compare the torsion angle label accuracy of the structural profiles on positions that are aligned to templates only. For this purpose, we implemented the profile methods summarized in Table 4.

Table 4. The implemented structural profile methods and their descriptions. Further details can be found in Sect. 2.4.

Structural profile and description
SP1: Eq. (1), Laplacian
SP2: Eq. (6), $a = 1$
SP3: Eq. (6), $a = 3$
SP4: Eq. (6), a varies wrt similarity interval
SP5: Eq. (7), quality, $a = 3$, (Pollastri et al.)
SP6: Eq. (8), E-value, $a = 3$
SP7: Eq. (10), E-value, align. score, $a = 3$
SP8: Eq. (11), E-value, norm. align. score, $a = 3$
SP9: Eq. (12), E-value, norm. align. score, BLOSUM62, $a = 3$
SP10: Eq. (13), E-value, align. score, a varies, BLOSUM62 scores if target is in High interval only
SP11: SP10 with windowing in (14)

In SP4, we modify the power of the sequence identity score term (*i.e.*, a) according to the similarity interval the target belongs to. For this purpose, we use the following mapping to define a:

$$a = \begin{cases} 1 \text{ if target in Low Interval} \\ 3 \text{ if target in Medium-Low Interval} \\ 5 \text{ if target in Medium Interval} \\ 7 \text{ if target in Medium-High Interval} \\ 9 \text{ if target in High Interval} \end{cases} \tag{18}$$

where the interval of the target is defined according to Table 2. In SP9 and SP10 we use the BLOSUM62 matrix to scale individual columns of the alignment as formulated in (12) and (13). In SP9, we employed the BLOSUM scores uniformly for all columns of the alignment whereas in SP10, we utilized the BLOSUM scores for targets in the "High" interval only. If the target belongs to one of the remaining four intervals then we turn off this score term in (13). In that case the equation takes the following form

$$\theta(t,i) = \begin{cases} s\, w_e\, I^a & \text{if } t = i \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

Once we compute a structural profile, we predict the torsion angle class of the aligned target residues by selecting the particular label that yields the maximum value in the corresponding column of the profile.

Following these definitions, the torsion angle class prediction accuracy of the structural profiles listed in Table 4 is summarized in Table 5 below. In this table, Overall 1 is the number of correctly predicted amino acids divided by the total number of amino acids for which a structural profile column is computed (*i.e.*, those that are aligned to at least one template). The second up to the sixth columns show accuracies for the five similarity intervals and are computed on each subset of the test set. Overall 2 is the average of the five accuracies obtained for Low to High intervals. It estimates the accuracy we would obtain had we used equal number of amino acids for each of the five intervals. Based on these results, the most accurate structural profile method is SP10 though other methods such as SP7, SP4 and SP8 are also quite effective. SP10 outperforms SP1, (the Laplacian count method), by 14.53 %, which is a statistically significant improvement. It is also better than SP5 by 1.2 %, which was proposed in [14,24]. This improvement is also statistically significant (with a p-value < 0.0001 from a two-tailed Z-test at a significance level of 0.01).

3.2 Combining Structural Profiles with Ab-Initio Predictions

After establishing that the new structural profile methods contain more accurate torsion angle information than the approaches proposed in the literature, we evaluated the accuracy when the structural profiles are combined with an ab-initio predictor. For this purpose, we trained our ab-initio method using the training set described in Sect. 2.7 and computed 7-state torsion angle predictions

Table 5. 7-state torsion angle prediction accuracy of structural profiles. L: Low, ML: Medium-Low, M: Medium, MH: Medium-High, H: High. Only target residues that are aligned to at least one template are considered.

Profile	L	ML	M	MH	H	Overall 1	Overall 2
SP1	66.49	69.83	71.94	74.46	75.66	74.17	71.68
SP2	66.85	70.71	73.93	80.07	81.96	79.18	74.70
SP3	66.68	72.20	77.82	86.64	92.87	87.64	79.24
SP4	66.85	72.20	79.40	87.92	93.47	88.30	79.97
SP5	66.53	72.13	77.28	87.07	92.73	87.50	79.15
SP6	67.15	73.26	78.89	87.13	93.06	88.03	79.90
SP7	66.48	74.11	80.23	87.66	93.30	88.40	80.36
SP8	66.59	73.91	80.13	87.31	93.25	88.33	80.24
SP9	65.05	72.19	78.30	86.39	93.68	88.14	79.12
SP10	66.55	74.11	80.71	87.36	93.69	88.70	80.48

on all the amino acids of the test set. We then combined those predictions with a structural profile as in (16). For the target residues that were not aligned to any template, we simply took predictions from the ab-initio method. Regarding the λ parameter (*i.e.*, the weight of the structural profile) we considered two possibilities. The first approach sets λ to 0.5 and the second one modifies it according to the similarity interval of the target according to the following function

$$\lambda = \begin{cases} 0.5 \text{ if target in Low Interval} \\ 0.6 \text{ if target in Medium-Low Interval} \\ 0.7 \text{ if target in Medium Interval} \\ 0.8 \text{ if target in Medium-High Interval} \\ 0.9 \text{ if target in High Interval} \end{cases} \tag{20}$$

In this equation, λ is gradually increased as the similarity interval of the target approaches to the "High" interval thereby giving more weight to the structural profile than the ab-initio predictor. Table 6 summarizes the accuracy of committee predictors that combine the ab-initio method with various structural profiles. In addition to the overall accuracy measure, we also included the segment overlap (SOV) measure that is used in 1D structure prediction to assess the accuracy at the segmental level [29]. The SOV measure depicts how well the predicted torsion label segments match the true segments and is biologically more meaningful than the residue level accuracy.

According to this table, the ab-initio+SP10 method (with variable λ parameter) is better than the ab-initio+SP5 method in all categories. The improvements are 2.78 % in Overall 1, 3.40 % in Overall 2, 3.57 % in SOV, 1.65 % in Low interval, 2.74 % in Medium-Low interval, 4.90 % in Medium interval, 4.82 % in Medium-High interval and 2.86 % in High interval. When ab-initio+SP10 is compared with ab-initio+SP1 (*i.e.*, the Laplacian method) for $\lambda = 0.5$,

Table 6. 7-state torsion angle prediction accuracy of methods that incorporate structural profiles with ab-initio predictions. The accuracy measures are computed on the first test set with 994 proteins. L: Low, ML: Medium-Low, M: Medium, MH: Medium-High, H: High. All target residues in the test set are considered.

Method	L	ML	M	MH	H	Overall 1	Overall 2	SOV
Ab-initio	72.36	73.96	73.58	73.35	74.01	73.83	73.45	71.33
Ab-initio + SP1, $\lambda = 0.5$	74.47	75.42	76.10	76.94	77.96	77.28	76.18	74.49
Ab-initio + SP5, $\lambda = 0.5$	72.39	74.19	74.19	75.53	78.28	76.99	74.92	74.52
Ab-initio + SP10, $\lambda = 0.5$	74.59	78.01	80.69	87.37	93.62	89.10	82.86	88.11
Ab-initio + SP5, λ as in (20)	72.94	75.21	77.22	83.86	90.99	86.70	80.04	85.19
Ab-initio + SP10, λ as in (20)	74.59	77.95	82.12	88.68	93.85	89.48	83.44	88.76

the improvements are 12.20 % in Overall 1, 6.68 % in Overall 2, 13.62 % in SOV, 0.12 % in Low interval, 2.59 % for Medium-Low interval, 4.59 % in Medium interval, 10.43 % in Medium-High interval and 15.66 % in High interval. Adjusting the λ parameter with respect to the similarity interval was particularly useful for the ab-initio+SP5 method. In other words, when λ is set to 0.5 uniformly for all similarity intervals, the accuracy of ab-initio+SP5 dropped significantly higher than the ab-initio+SP10. This shows that the proposed structural profile SP10 is more useful than SP5 when combined with the ab-initio method. This is because torsion label errors of SP10 and the ab-initio method overlap less as compared to SP5 and therefore SP10 provides a better complement to the ab-initio predictor. Another observation one can make is the improvement over the ab-initio method when structural profiles are incorporated. This is true even for the Low interval (an improvement of 2.23 %) and is partly because of the sensitive nature of HMM-HMM profile alignments and also because HHsearch uses predicted secondary structure from PSIPRED [11] to align a pair of HMMs.

3.3 Optimizing Parameters

We optimized the parameters a, W and λ on the validation set as explained in Sect. 2.6. We observed that taking different values for the window parameter W did not alter the accuracy of predictions considerably. Therefore this parameter is selected as nearly constant (Table 7).

Table 8 compares the accuracy of 7-state torsion angle prediction of various methods on the second test set of 497 proteins. In this table, except for the last row all the hits that score above the threshold are used during profile construction (i.e. no windowing is applied). Based on these results, optimizing the parameters of the profile models improves the accuracy of predictions as compared to the case where a and λ are fixed. However selecting the optimums (see the sixth row of Table 8) was not considerably better than the results in row five. The reason is because the cost curve that we are trying to optimize is not strongly convex (i.e. nearly flat around the maxima) and these are merely the variations caused by using different datasets for optimization and accuracy estimation.

Table 7. Optimum values for λ, W, and a. The values that give the best amino acid level accuracy on the validation set are selected. The optimization is performed separately for each HHsearch similarity interval. The improvements are computed with respect to the accuracy of *ab-initio* predictions.

Interval	L	ML	M	MH	H
λ^*	0.5	0.6	0.6	0.6	0.7
a^*	2	3	4	7	7
W^*	20.0	10.0	10.0	10.0	10.0

Table 8. 7-state torsion angle prediction accuracy of methods that incorporate structural profiles with ab-initio predictions. The accuracy measures are computed on the second test set with 497 proteins. L: Low, ML: Medium-Low, M: Medium, MH: Medium-High, H: High. All target residues in the test set are considered.

Method	L	ML	M	MH	H	Overall 1	Overall 2	SOV
Ab-initio	71.77	74.67	73.04	74.05	74.18	73.97	73.54	71.49
Ab-initio + SP1, $\lambda = 0.5$	73.29	76.08	75.29	77.70	78.19	77.41	76.27	74.46
Ab-initio + SP10, $a = 3$, $\lambda = 0.5$	73.29	77.87	80.82	87.24	93.24	89.11	82.49	87.97
Ab-initio + SP10, a as (18), $\lambda = 0.5$	73.44	77.87	81.15	87.99	93.61	89.20	82.81	88.09
Ab-initio + SP10, a as (18), λ as (20)	73.44	77.74	82.52	89.03	93.89	89.60	83.32	88.77
Ab-initio + SP11, W, λ, a optimized	73.31	77.58	82.23	89.02	93.97	89.33	83.22	88.34

3.4 Other Approaches

In addition to the structural profile methods described above, we also considered three other approaches. The first one incorporates the probability score of HMM-profile alignments into (13) as a multiplicative factor to globally scale the templates and the second method incorporates the column score of HHsearch alignments to amplify local regions that are well conserved (e.g. motifs). These two approaches did not bring any reasonable change in the accuracy measures (result not shown). As a third approach we considered employing Henikoff weights [9] to scale the count information of templates before constructing the structural profiles. We applied this weighting procedure for the following three scenarios: (1) weights based on matched residues only, (2) weights based on torsion angle labels only, (3) weights based on residue and torsion angle tuples. Unfortunately, in all three cases, torsion angle prediction accuracy was significantly lower than the level achieved by other scaling methods considered in this paper (result not shown). Note that we did not consider utilizing a background distribution in (4) for torsion angle labels mainly because we use predictions from our ab-initio method, which would eventually contain a more accurate torsion angle representation than a simple background distribution.

Finally, we would like to state that we are unable to compare our torsion angle class predictor with the literature mainly because there is no other work that performs 7-state torsion angle prediction on the same set of alphabet. However we had shown in an earlier paper that a 5-state version of our predictor provides results comparable to the state-of-the-art in the ab-initio setting [3].

4 Conclusion

In this paper, we propose novel methods for scaling templates to construct structural profiles of torsion angle states. Though we use the score terms of the HHsearch method, our approach is generic and most of the structural profile methods proposed in this work can also be implemented using other alignment methods including PSI-BLAST. Second, the scaling techniques can be applied in many other tasks such as secondary structure prediction, solvent accessibility prediction, contact map prediction, and 3D structure prediction. Third, they can easily be incorporated into other methods that have been developed for deriving structural profiles from templates.

The proposed methods can be improved in several ways. First of all, the templates can be scaled in a position-specific manner using the confidence scores, which are now available in HHBlits (the new version of HHsearch). Second, instead of using a linear model, the ab-initio predictions can be combined with structural profiles using more advanced models such as neural networks. We expect that all these approaches will potentially improve the accuracy of structure prediction tasks further.

References

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997)
2. Aydin, Z., Singh, A., Bilmes, J., Noble, W.S.: Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. BMC Bioinform. **12**, 154 (2011)
3. Aydin, Z., Thompson, J., Bilmes, J., Baker, D., Noble, W. S.: Protein torsion angle class prediction by a hybrid architecture of bayesian and neural networks. In: 13th International Conference on Bioinformatics and Computational Biology (2012)
4. Berjanskii, M.V., Neal, S., Wishart, D.S.: PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res. **34**, W63–W69 (2006). (Web Server Issue)
5. Blum, B., Jordan, M., Kim, D., Das, R., Bradley, P., Baker, D.: Feature selection methods for improving protein structure prediction with Rosetta. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems 20, pp. 137–144. MIT Press, Cambridge (2008)
6. Cheng, J., Tegge, A.N., Baldi, P.: Machine learning methods for protein structure prediction. IEEE Rev. Biomed. Eng. **1**, 41–49 (2008)
7. Cong, P., Li, D., Wang, Z., Tang, S., Li, T.: Spssm8: an accurate approach for predicting eight-state secondary structures of proteins. Biochimie **95**(12), 2460–2464 (2013)
8. Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., Zhou, Y.: SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. PLoS One **7**(2), e30361 (2012)
9. Henikoff, S., Henikoff, J.G.: Position-based sequence weights. J. Mol. Biol. **243**, 574–578 (1994)

10. Hobohm, U., Sander, C.: Enlarged representative set of protein structures. Protein Sci. **3**, 522–524 (1994)
11. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**, 195–202 (1999)
12. Li, D., Li, T., Cong, P., Xong, W., Sun, J.: A novel structural position-specific scoring matrix for the prediction of protein secondary structures. Bioinformatics **28**(1), 32–39 (2012)
13. Mooney, C., Pollastri, G.: Beyond the twilight zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins Struct. Funct. Bioinform. **77**, 181–190 (2009)
14. Pollastri, G., Martin, A.J.M., Mooney, C., Vullo, A.: Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinform. **8**, 201 (2007)
15. Rangwala, H., Karypis, G.: Introduction to Protein Structure Prediction: Methods and Algorithms. Wiley, Hoboken (2011)
16. Remmert, M., Biegert, A., Hauser, A., Soding, J.: Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nat. Meth. **9**(2), 173–175 (2011)
17. Shen, Y., Delaglio, F., Cornilescu, G., Bax, A.: TALOS+: a hybrid method for predicting protein backbone torsion angles from nmr chemical shifts. J. Biomol. NMR **44**(4), 213–223 (2009)
18. Singh, H., Singh, S., Raghava, G.P.S.: Evaluation of protein dihedral angle prediction methods. PLoS One **9**(8), e105667 (2014)
19. Soding, J.: Protein homology detection by HMM-HMM comparison. Bioinformatics **21**, 951–960 (2005)
20. Soding, J.: Quick guide to HHsearch (2006). ftp://toolkit.genzentrum.lmu.de/pub/HHsearch/old/HHsearch/HHsearch1.5.1/HHsearch-guide.pdf
21. Soding, J., Remmert, M., Hauser, A.: HH-suite for sensitive sequence searching based on HMM-HMM alignment (2012). ftp://toolkit.genzentrum.lmu.de/pub/HH-suite/hhsuite-userguide.pdf
22. Song, J., Tan, H., Wang, M., Webb, G.I., Akutsu, T.: TANGLE: two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences. PLoS One **7**(2), e30361 (2012)
23. Sun, J., Tang, S., Xiong, W., Cong, P., Li, T.: Dsp: a protein shape string and its profile prediction server. Nucleic Acids Res. **40**(W1), W298–W302 (2012)
24. Walsh, I., Bau, D., Martin, A.J.M., Mooney, C., Vullo, A., Pollastri, G.: Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Struct. Biol. **9**, 5 (2009)
25. Wang, G., Dunbrack Jr., R.L.: PISCES: a protein sequence culling server. Bioinformatics **19**, 1589–1591 (2003). http://dunbrack.fccc.edu/PISCES.php
26. Wang, G., Dunbrack Jr., R.L.: PISCES: recent improvements to a pdb sequence culling server. Nucleic Acids Res. **33**, W94–W98 (2005)
27. Wu, S., Zhang, Y.: ANGLOR: A composite machine-learning algorithm for protein backbone torsion angle prediction. PLoS One **3**(10), e3400 (2008)
28. Wu, S., Zhang, Y.: MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins Struct. Funct. Bioinform. **72**(2), 547–556 (2008)
29. Zemla, A., Venclovas, C., Fidelis, K., Rost, B.: A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins **34**, 220–223 (1999)
30. Zhou, Y., Duan, Y., Yang, Y., Faraggi, E., Lei, H.: Trends in template/fragment-free protein structure prediction. Theo. Chem. Acc. **128**, 3–16 (2011)