Developing Structural Profile Matrices for Protein Secondary Structure and Solvent Accessibility Prediction

Zafer Aydin [1,*], Nuh Azginoglu [2], Halil Ibrahim Bilgin [1], and Mete Celik [3]

[1]Department of Computer Engineering, Abdullah Gul University, Kayseri, 38080, Turkey
[2]Department of Computer Engineering, Nevsehir Haci Bektas Veli University, Nevsehir, 50300, Turkey
[3]Department of Computer Engineering, Erciyes University, Kayseri, 38039, Turkey.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Predicting secondary structure and solvent accessibility of proteins are among the essential steps that preclude more elaborate 3D structure prediction tasks. Incorporating class label information contained in templates with known structures has the potential to improve the accuracy of prediction methods. Building a structural profile matrix is one such technique that provides a distribution for class labels at each amino acid position of the target.

Results: In this paper, a new structural profiling technique is proposed that is based on deriving PFAM families and is combined with an existing approach. Cross-validation experiments on two benchmark datasets and at various similarity intervals demonstrate that the proposed profiling strategy performs significantly better than Homolpro, a state-of-the-art method for incorporating template information, as assessed by statistical hypothesis tests.

Availability: The DSPRED method can be accessed by visiting the PSP server at http://psp.agu.edu.tr. Source code and binaries are freely available at https://github.com/yusufzaferaydin/dspred.

Contact: zafer.aydin@agu.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins are essential molecules for the biological processes in cells of the living beings. Knowing the structure of a protein is important as it helps to understand its functional activity. Furthermore, it is useful for drug design studies that model ligand protein interactions. Because experimental determination of protein structure is labor intensive and may take weeks researchers have developed an interest in alternative solutions such as computational prediction of three-dimensional (3D) structure.

One of the preliminary steps of 3D structure prediction is to estimate one or two dimensional structural attributes such as secondary structure, solvent accessibility, torsion angles, and contact maps, which are subsequently employed as inputs to more elaborate energy minimization algorithms (Yang and Zhang, 2016). Several methods have been proposed to predict these properties most of which are based on training machine learning models. The input features of these prediction methods typically include sequence profiles in the form of position-specific scoring matrices (PSSMs) and structural profile matrices both of which are derived by aligning the target protein to proteins (or to profile models that represent proteins) in a database of interest. Though there has been a lot of interest in developing sophisticated prediction models less effort has been made by the researchers on developing better feature parameters. However, it is well-known in machine learning literature that extracting more informative and discriminative features contributes positively on the prediction accuracy.

Developing new feature extraction methods for predicting protein structure may consider two main directions: deriving better sequence profiles or structural profiles. Methods developed for the first objective typically align the target protein with amino acid sequences in a large sequence database to estimate position-specific sequence profiles that represent the statistical propensity of observing each of the twenty amino

acids at individual positions of the target. In this category structure information of the proteins are not employed explicitly. On the other hand, methods in the second category focus on finding template proteins with known structures and summarize the label frequency information of the templates in the form of structural profile matrices, which can be more effective for predicting structure of proteins. A structural profile is a collection of discrete probability distributions each showing the propensity of a target amino acid to be in one of the structure states (e.g. in secondary structure prediction the state space can be {H,E,L} and in solvent accessibility {e,b}). Employing a structural profile matrix is effective also because of the following. Although the protein sequence databases contain hundreds of millions of proteins and the Protein Data Bank (PDB) contains hundreds of thousands of solved structures, it is estimated that the number of distint structures a protein can fold into is on the order of thousands. Depending on the similarity between target and templates class label information can be potentially more useful than the information contained in a sequence profile.

To date, structural profile matrices have been employed in various methods to improve the accuracy of machine learning models developed for predicting structural properties of proteins. Pollastri et al. used structural profile matrices in the input feature vector of bi-directional recurrent neural networks and obtained improvements in secondary structure and solvent accessibility prediction (Pollastri *et al.*, 2007). Mooney and Pollastri employed predicted structural features to align the target with the templates and to derive a structural profile, which is fed as input to recurrent neural networks with sequence-based profile features (Mooney and Pollastri, 2009). Cheng et al. followed a data mining approach that searches for fragment structures in PDB to improve prediction accuracy of GOR V method (Cheng *et al.*, 2007). A similar approach is proposed in Lin et al., which uses match rates of short fragments with known structure aligned to target in combination with PSIPRED (Lin *et al.*, 2005). Walsh et al. incorporated template information to improve the accuracy of two-dimensional distance map prediction (Walsh *et al.*, 2009). Zhang et al. developed torsion angle and solvent accessibility profiles for protein fold recognition (Zhang *et al.*, 2008). Magnan and Baldi employed Homolpro as a post-processing module after bi-directional recurrent neural networks model (1D-BRNN) to incorporate structural label information from templates into secondary structure and solvent accessibility prediction (Magnan and Baldi, 2014). Li et. al introduced a secondary structure prediction method called SPSSMPred, which employs features from a sequence-based PSSM as well as a structural profile matrix (Li *et al.*, 2012). Recently Zhou et al. introduced a template library called SIPSS, which is periodically updated as new structures are deposited into PDB (Zhou *et al.*, 2017). The methodology used in this work is similar to the paper by Li et al. In another recent paper, Aydin et al. developed structural profile matrices by aligning the target with PDB proteins using HHblits (Remmert *et al.*, 2012) and obtained improvements in torsion angle class prediction (Aydin *et al.*, 2015).

Most of the related work in the literature do not apply weights on templates when constructing structural profile matrices. It has been shown in Aydin et al. that such type of weighting can improve the accuracy of prediction considerably (Aydin *et al.*, 2015). However this technique has not been applied to secondary structure and solvent accessibility prediction yet. Furthermore there is no work that employs sequence based motifs such as PFAM motifs (Finn *et al.*, 2016), which might be able to capture sequence-based patterns related to structural and functional context of proteins.

This paper introduces a new technique for computing a structural profile matrix. Starting from the target, it performs a PFAM domain family search by PfamScan and extracts templates from PDB that belong to the same PFAM family as the target. In the next step, the templates are aligned with the target using T-Coffee and blastp programs. Finally, weighted frequency of occurrence counts are computed to derive the structural profile matrix. The proposed method is combined with the structural profiling approach developed earlier (Aydin *et al.*, 2015) for torsion angle class prediction and is incorporated into DSPRED, which is a two-stage hybrid classifier developed for predicting one-dimensional structure of proteins including secondary structure, solvent accessibility and torsion angle class information.

2 Methods

2.1 Problem Definition

One-dimensional protein structure prediction aims to assign a structural label to each amino acid of a given protein. For example, in secondary structure class prediction, the goal is to assign one of the labels H, E, or L to each amino acid. Detailed definition of prediction tasks can be found in Supplementary Section S1.

2.2 Datasets

This section describes the datasets used to compute statistics about structural profile matrices and to evaluate the accuracy of DSPRED in secondary structure and solvent accessibility prediction.

2.2.1 NRNPDB992

NRNPDB992 dataset contains 992 proteins from the non-redundant (NR) database of NCBI with unknown structures. It is constructed by first selecting a set of 1000 proteins randomly from the NR database and then eliminating those proteins that have known structures in PDB. NRNPDB992 is used to obtain the percentage of targets that are matched with at least one template so that a structural profile matrix can be computed and the percentage of target amino acids to which at least one template residue is aligned.

2.2.2 CB513

CB513 is one of the established and difficult benchmarks used to assess the prediction accuracy of one-dimensional structure prediction methods (Cuff and Barton, 1999). It contains 513 chains and 84119 amino acids. In this paper, CB513 is used to evaluate the secondary structure prediction accuracy of DSPRED method and various structural profiling techniques. This dataset can be downloaded from Jpred's distribution material website http://www.compbio.dundee.ac.uk/jpred/legacy/data/.

2.2.3 EVAset

EVAset is another benchmark that contains proteins from PDB and is designed for evaluating the accuracy of prediction methods in structure prediction tasks (Koh *et al.*, 2003). The original dataset contains 3074 proteins. After removing proteins shorter than 30 amino acids there remained a set of 2876 targets with a total of 584595 amino acids. In this paper, EVAset is used to evaluate the accuracy of DSPRED method in secondary structure and solvent accessibility prediction tasks.

2.2.4 PDB99

This is the database of HMM-profiles used in the second step of HHblits. To derive this databse, first, a non-redundant set of PDB proteins were downloaded using the software from the PISCES server (http://dunbrack.fccc.edu/PISCES.php) (Wang and Dunbrack, Jr., 2003) by setting the threshold to 99%, which eliminates all protein pairs that have percentage of identity score above this threshold. This step produced a set of 23936 proteins. In the next step, HMM-profiles are built by following the steps in Section 3.5 "Building customized databases" of the HHsuite

user guide (https://github.com/soedinglab/hh-suite/blob/master/hhsuite-userguide.pdf).

2.3 Assigning structure labels

The secondary structure label information for the CB513 dataset was originally downloaded from the Jpred's distribution material website along with the amino acid sequences. The secondary structure and solvent accessibility labels of proteins in EVAset are obtained by first downloading the pdb files of the targets using the get_pdb.py script of Rosetta (https://www.rosettacommons.org/software) and then by running the DSSP program, which processes the 3D coordinate information in pdb files. The output of DSSP includes an 8-state label sequence for secondary structure representation and a sequence of solvent accessibility scores, which include an accessibile surface area for each amino acid (i.e. the absolute accessibility). In this work, the 8-state secondary structure label sequence is converted to 3-state using the following transformation: $\{H, G, I\} \rightarrow H$; $\{E, B\} \rightarrow E$; $\{\text{' '}, S, T\} \rightarrow L$. The absolute solvent accessibility scores are converted to relative accessibility by dividing the accessible surface area of each amino acid by its maximum accessibility score. Detailed explanation on this transformation can be found in https://en.wikipedia.org/wiki/Relative_accessible_surface_area. The relative accessibility values are then compared to one or more thresholds and transformed to discrete labels depending on which interval they fall into. In this paper, a single threshold is used, which is set to 25%. If the relative accessibility of an amino acid is greater than 25% then it is assigned to the "e" (i.e. exposed) class. Otherwise its label becomes "b" (i.e. buried).

2.4 Feature Extraction

This section explains how the input features of the prediction method are computed.

2.4.1 PSI-BLAST PSSM features

Proteins in CB513 and EVAset benchmarks are aligned with proteins in the NCBI's NR database using PSI-BLAST (see protein BLAST at https://blast.ncbi.nlm.nih.gov/Blast.cgi) with the following parameter settings: e-value=10, number of iterations=3, and inclusion e-value=0.001. A position-specific scoring matrix (PSSM) of size $N \times 20$ is computed for each protein using the out_ascii_pssm option of PSI-BLAST, where N is the number of amino acids in the target. The PSSM values are scaled to interval [0,1] as in in Aydin et al. (Aydin *et al.*, 2011) by applying a sigmoidal transformation individually to each parameter and are used as input features of the DSPRED method. Details on PSSM computation using PSI-BLAST can be found in the paper by Altschul et al. (Altschul *et al.*, 1997).

2.4.2 HHMAKE PSSM features

To derive HHMAKE PSSM features, proteins in CB513 and EVAset are aligned with the NR20 database (a reduced version of NCBI's NR database) using the first step of HHblits (https://toolkit.tuebingen.mpg.de/#/tools/hhblits) by setting the number of iterations to 2 and all other parameters to their default settings. This step produces a multiple alignment between the target and hit proteins. In the next step, an HMM-profile model is constructed using the hhmake utility of HHblits, which contains match, insertion, and deletion states as well as a background amino acid score table. Each match state and the background table contains a set of 20 scores, one for each amino acid. Since the raw output of hhmake includes scores in $-1000 * \log(value)$ format, the scores in match states and the background scores are first divided by 1000 and then PSSM values are computed by subtracting each match score from the corresponding background score. Finally, similar to PSI-BLAST features, the PSSM values are normalized to interval [0,1] by a sigmoidal transformation and used as input features of DSPRED (Aydin *et al.*, 2011).

2.4.3 Generating structural profile matrices using HHblits: SP1

A structural profile matrix represents the propensity of each amino acid of the target to be in one of the structural class states in the form of discrete probability distributions. The dimension of this matrix is $N \times K$ where N is the number of amino acids in target, $K = 3$ for secondary structure prediction and $K = 2$ for solvent accessibility prediction. Since each row is a discrete probability distribution, the sum of the scores in each row should be 1. Detailed definition of a structural profile matrix can be found in Supplementary Section S2.

The first type of structural profile matrix, denoted as SP1, is computed using the HHblits method (https://toolkit.tuebingen.mpg.de/#/tools/hhblits). For this purpose, the target is initially aligned with proteins in the NR20 sequence database using the hhblits and hhmake utilities of HHblits as explained in Section 2.4.2. This step produces an HMM-profile model for the target, which is then aligned with the HMM-profiles in the PDB99 database by setting the number of iterations to 1 and all the other parameters to their default settings. In the last step, a position-specific structural profile matrix is computed using the weighted frequency information of the structure labels assigned from PDB99 to amino acids of the target (Aydin *et al.*, 2015). This is formulated as follows

$$S(i,j) = \begin{cases} \frac{C(i,j)}{\sum_j C(i,j)} & \text{if } |A(i)| > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where S is the normalized count matrix representing the structural profile matrix, i is the index for amino acids of the target such that $1 \leq i \leq N$ with N representing the number of amino acids in target, j is the index for structure labels such that $1 \leq j \leq K$ with $K = 3$ for secondary structure prediction and $K = 2$ for solvent accessibility prediction, $C(i,j)$ is the unnormalized count matrix, $A(i)$ is the set of template residues aligned to the i^{th} residue of the target, and $|A(i)|$ is the number of residues in $A(i)$. The above equation splits the set of target resides into two groups: those that are aligned to at least one template residue (i.e. $|A(i)| > 0$) and those that are not aligned to any template residue (i.e. $|A(i)| = 0$). The second set of residues can occur because HHblits computes local alignments, which may also include gapped regions. The unnormalized count matrix is computed as

$$C(i,j) = \sum_{R(i,k)} \theta(t(i,k), j), \tag{2}$$

where C is the count matrix, $R(i,k)$ is the amino acid residue of the k^{th} database protein aligned to the i^{th} position of the target, $t(i,k)$ is the true class label of $R(i,k)$, and θ is the occurrence count function expressed as

$$\theta(t(i,k), j) = \begin{cases} s\, w_e\, I^a\, c(i) & \text{if } t(i,k) = j \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where s is the raw score of the alignment, w_e is the e-value weight defined in Aydin et al. (Aydin *et al.*, 2015), I is the sequence identity score of the alignment between the target and the k^{th} template, a is the integer power factor, and $c(i)$ is the confidence score assigned by HHblits to the i^{th} residue of the target. In this equation, the integer a amplifies the contribution of structurally close templates preventing them to be swamped with many structurally distant templates or false positives. In the present study, the following values are chosen for this parameter: where PC is the threshold for the maximum allowed percentage of sequence identity score of the target-template alignments used to construct the structural profile matrix. Different from Aydin et al. (Aydin *et al.*, 2015) Eq. 3 includes the confidence score $c(i)$, which is a position-specific weight term

Table 1. The a parameter with respect to the maximum allowed percentage of sequence identity score for a target-template pair

PC	20	30	40	50	60	70	80	90	100
a	1	2	3	4	5	6	7	8	9

that participates in the computation of weighted frequency of occurrence counts.

2.4.4 Generating structural profile matrices using PfamScan: SP2

PFAM is a database containing protein families each represented by multiple sequence alignments and hidden Markov models (https://pfam.xfam.org) (Finn *et al.*, 2016). It offers a search tool called PfamScan, which can be used to find the domain family of a given protein. Starting from the amino acid sequence of the query, PfamScan computes alignments against a library of HMM-profiles using HMMER3 (http://hmmer.org) and reports the family (or families) with the best alignment score(s).

In this paper, to derive the second structural profile matrix (i.e. SP2), first, the domain family of the target is found by running PfamScan in default settings. Then, other proteins belonging to the same family are detected by searching the pdb map file of the PFAM database for the PFAM domain ID. In the next step, templates belonging to the same PFAM family are aligned pairwise with the target using blastp program of NCBI's BLAST software (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the percentage of sequence identity scores are recorded. To improve the alignment quality of blastp, the target and the templates are aligned by the T-Coffee multiple alignment software (http://tcoffee.crg.cat) and the structural profile matrix is computed based on the residue matches produced by T-Coffee and using Eqs. 1, 2, and 4, which is given below

$$\theta(t(i,k),j) = \begin{cases} I^a & \text{if } t(i,k) = j \\ 0 & \text{otherwise} \end{cases} . \quad (4)$$

Note that Eq. 4 is a simplified version of Eq. 3 and includes merely the power of the sequence identity score from the blastp alignment.

2.4.5 Eliminating templates at different similarity thresholds

To measure the contribution of the template similarity on the prediction accuracy, templates that score above the percentage of sequence identity threshold denoted as PC are excluded from each structural profile matrix. For this purpose, the following 9 thresholds are considered: 20, 30, 40, 50, 60, 70, 80, 90, and 100. Then for each threshold, structural profile matrices are computed for all targets in CB513 and EVAset by eliminating those templates that have percentage of sequence identity score greater than this maximum score threshold. Since all proteins in CB513 and EVAset are PDB proteins, templates that share the same PDB ID as the target (including those that have the same PDB ID and same chain ID as well as those that have the same PDB ID but different chain ID) are always excluded and not used for estimating the structural profile matrices.

2.5 DSPRED Prediction Method

In this paper, the secondary structure and solvent accessibility information is predicted using the DSPRED method, which is a two-stage hybrid classifier developed for estimating one-dimensional structural attributes of proteins (Aydin *et al.*, 2011). The first stage contains dynamic Bayesian network (DBN) classifiers and the second stage employs a support vector machine (SVM) classifier. Different DBN models are trained for PSI-BLAST PSSM and HHMAKE PSSM features, respectively and each DBN computes a probability distribution of class labels given input PSSM features. These are denoted as Distribution 1 and Distribution 2, respectively for PSI-BLAST PSSM and HHMAKE PSSM, which have the same form as a structural profile matrix. Then these two distributions and the structural profile matrices are averaged to obtain Distribution 3 as explained in Section 2.5.1. Details of DSPRED can be found in Supplementary Section S3.

2.5.1 Weighted average of DBN outputs and structural profiles

As explained in Section 2.5 and in Supplementary Section S3, Distribution 3 (one of the feature subsets of the SVM model) is computed as the weighted average of DBN outputs and the structural profile matrices. This is achieved by

$$D_3(i,j) = w_1 D_1(i,j) + w_2 D_2(i,j) + w_3 S_1(i,j) + w_4 S_2(i,j) \quad (5)$$

where D_1, D_2, and D_3 are Distributions 1 to 3, respectively, S_1 and S_2 denote structural profile matrices SP1 and SP2 explained in Sections 2.4.3 and , respectively, j is the index for class label, i is the index for amino acid position, w_1, w_2, w_3, and w_4 represent the weights of Distribution 1, Distribution 2, SP1, and SP2, respectively. The dimensions of D_1, D_2, D_3, S_1, and S_2 are N by K where N is the number of amino acids in target, $K = 3$ for secondary structure prediction and $K = 2$ for solvent accessibility prediction. These weight terms satisfy $w_1 + w_2 + w_3 + w_4 = 1$ and each weight is applied uniformly to all amino acid positions of the corresponding distribution. However, if there is no aligned residue to a target amino acid in SP1 and/or SP2 then the the corresponding weight term(s) are set to zero at that amino acid only. For instance, if there is no aligned residue to a target amino acid in HHblits alignments (i.e. for SP1) but alignments are available in blastp (i.e. for SP2) then w_3 is set to zero for that amino acid only. Similarly, if there is no aligned residue to a target amino acid both in HHblits and blastp alignments then w_3 and w_4 are set to zero for that amino acid only. This allows Distributions 1 and 2 to serve as background models in positions where there is no template residue aligned to those positions due to local nature of the alignments computed by HHblits and blastp. The following senarios are implemented to select weights w_1, w_2, w_3, and w_4:

No SP implements the case that only uses DBN outputs and excludes structural profile matrices from consideration. In this scheme, $w_1 = 0.5$, $w_2 = 0.5$, $w_3 = 0.0$, and $w_4 = 0.0$, which is uniformly applied to all amino acid positions.

SP1 only implements the scenario that combines outputs of DBN models as well as the first structural profile matrix (SP1) excluding SP2. This scheme sets $w_1 = 1/3$, $w_2 = 1/3$, $w_3 = 1/3$, and $w_4 = 0.0$.

SP1+SP2 combines the outputs of DBNs and the two structural profile matrices by selecting $w_1 = 1/4$, $w_2 = 1/4$, $w_3 = 1/4$, and $w_4 = 1/4$.

2.6 Incorporating Homolpro into DSPRED

Homolpro is a software that is developed for transferring labels from PDB templates to predicted class label sequence of a target (Magnan and Baldi, 2014). It performs a blastp alignment against the pdb-full database of 139859 redundant proteins and constructs a structural profile matrix using Laplacian counts (i.e. frequency counts of 1) for each template residue matched to target. In the next step, if an amino acid on the target is aligned with a template residue, the predicted class label of that amino acid is

updated by copying the label that has the highest score at the corresponding column of the structural profile matrix. Homolpro requires the templates to satisfy the following constraints: (1) a minimum of 10 amino acids should be aligned with the target for secondary structure and 30 amino acids for solvent accessibility prediction, (2) the percentage of sequence identity score should be at least 45% for secondary structure and 70% for solvent accessibility prediction, (3) the BLAST e-value should be less than 10^{-9}, and (4) positive substitution score should be at least 55% for secondary structure and 75% for solvent accessibility prediction.

In this paper, the Homolpro software version 1.1 is used as a post-processing block (http://download.igb.uci.edu) after the DSPRED method. For this purpose, first, the predicted label sequence is computed using DSPRED by setting w_3 and w_4 to zero, which is sent as input to Homolpro to update the labels predicted by the SVM classifier. To realize the experimental settings of the present work, Homolpro is customized by: (1) setting the minimum alignment score to 0, (2) introducing a maximum alignment score (which corresponds to the PC threshold in this work that takes values from 20% to 100%), and (3) replacing the pdb-full database with PDB99 (proteins and the label sequences of PDB99 only). Introducing a maximum alignment score threshold allows to eliminate templates that score above the threshold and to analyze the contribution of template information at different similarity levels.

2.7 Accuracy Measures

The following accuracy measures are used: overall prediction accuracy, segment overlap measure, class-specific recall, precision, and Matthew's correlation coefficient measures. The overall accuracy is computed as the percentage of correctly predicted residues divided by the total number of amino acids. Segment overlap measure aims to compute the ratio of overlap between the true and predicted label segments. Detailed explanation of these metrics can be found in the paper by Aydin et al. (Aydin *et al.*, 2011).

3 Results and Discussion

3.1 Distribution of the sequence identity score in HHblits alignments

We analyzed the distribution of the sequence identity score between a target and proteins in the PDB99 dataset at the end of an HHblits alignment (see Supplementary Section S4). Based on this analysis, the majority of the templates found by HHblits have sequence identity scores in range 20%-40%.

3.2 Template Matching Statistics of SP1 and SP2

A natural point of interest is the potential of a structural profile matrix derivation method to find templates to a given target. To explore this, we computed the following statistics for SP1 and SP2 matrices: the percentage of targets that receive at least one hit from PDB and the percentage of amino acid positions that are aligned to a residue of a PDB template (see Supplementary Section S5 for details). Based on this analysis, it is possible to find templates using HHblits and/or PfamScan even for difficult cases that only have distant templates available.

3.3 The effect of employing a structural profile matrix

This section compares the prediction accuracy of DSPRED for the case that does not use any structure profile matrices (i.e. No SP) to the condition that uses the first structural profile matrix (i.e. SP1 only with $w_1 = w_2 = w_3 = 1/3$ and $w_4 = 0$) with PC threshold equal to 20 (employing the most distant templates only) and PC threshold equal to 100 (employing all templates available). Table 2 includes the overall accuracy, segment overlap measure of DSPRED as well as standard deviation of overall accuracy on CB513 and EVAset for secondary structure prediction. A 7-fold cross-validation experiment is performed on CB513 and a 10-fold cross-validation on EVAset. Standard deviation is computed using the accuracy obtained for each fold of the cross-validation. According to this table, employing SP1 profiles improves the overall accuracy of DSPRED by 1.14% on CB513 and 0.95% on EVAset when distant templates are used only. The improvements for the SOV measure are 1.05% on CB513 and 1.38% on EVAset. The 1.14% improvement in overall accuracy of secondary structure preciction on CB513 is statistically significant as assessed by a two-tailed Z-test for comparing proportions (https://onlinecourses.science.psu.edu/stat414/node/268/) with a Z-score of 6.0691 and a p-value of approximately 0 at a confidence level of 95%. Similarly, the 0.95% improvement in overall accuracy of secondary structure prediction on EVAset is statistically significant according to a two-tailed Z-test with a Z-score of 13.7922 and a p-value of approximately 0 at a confidence level of 95%. This demonstrates that employing structural profiles that are constructed even using only distant templates improves the accuracy of the prediction model significantly. The improvements are much higher for higher values of the PC threshold (i.e. when more similar templates are employed in computing the structural profiles). When PC threshold is set to 100 allowing all templates available, the accuracy of DSPRED improves by 10.03% on CB513 and 7.15% on EVAset for secondary structure prediction. The improvements for the SOV measure are 11.67% on CB513 and 7.48% on EVAset. Detailed accuracy metrics for these experiments can be found in Supplementary Tables S4 and S5 and Supplementary Table S6 includes results for solvent accessibility prediction on EVAset.

Table 2. The secondary structure prediction accuracies of DSPRED when no structural profile matrix is used and when SP1 is used with a PC threshold of 20 employing the most distant templates only and a threshold of 100 employing all the templates available.

Dataset	Task	Method	Q3	SOV	std
CB513	SS3	No SP	81.64	77.30	0.821
CB513	SS3	SP1 only, PC=20	82.78	78.35	0.897
CB513	SS3	SP1 only, PC=100	91.67	88.97	0.805
EVAset	SS3	No SP	82.89	78.08	0.314
EVAset	SS3	SP1 only, PC=20	83.84	79.30	0.323
EVAset	SS3	SP1 only, PC=100	89.64	85.57	0.450

3.4 Incorporating templates at different similarity levels

To analyze the effect of template similarity on the prediction accuracy, the PC threshold is increased from 20% to 100% with increments of 10 allowing only the templates having percentage of sequence identity scores lower than the threshold in constructing the structural profile matrix. Table 3 includes accuracy metrics from 7-fold cross-validation experiments performed on CB513 for protein secondary structure prediction, Table 4 contains accuracy metrics from 10-fold cross-validation experiments performed on EVAset for secondary structure prediction, and Table 5 demonstrates accuracy metrics from 10-fold cross-validation experiments performed on EVAset for solvent accessibility prediction. In Homolpro, SP1 only, and SP1+SP2 columns, numbers inside parantheses are segment overlap measures and those outside include the overall accuracies. HT1 column includes two-tailed Z-test results between the overall accuracy

metrics of SP1 only and Homolpro columns at a confidence level of 95% (i.e. a significance level of 0.05), in which the numbers inside paranthesis include p-values and those that are outside are the Z-scores. Similarly, HT2 column contains two-tailed Z-test results between the overall accuracy values of SP1+SP2 and Homolpro columns at the same confidence level as HT1.

According to the hypothesis test results, DSPRED with SP1 only performs significantly better than Homolpro on CB513 benchmark at PC thresholds of 20 to 70 (including templates with low to moderate similarity), comparable to Homolpro at PC values of 80 and 100 and slightly lower than Homolpro when PC is set to 90. SP1+SP2 case performs significantly better than Homolpro on CB513 at all PC thresholds except for PC equal to 90 at which the methods perform comparably (Table 3). In secondary structure and solvent accessibility prediction experiments on EVAset, both SP1 only and SP1+SP2 approaches perform significantly better than Homolpro at all PC thresholds (Tables 4 and 5). The improvements in overall accuracy are approximately 1%-2.4% on CB513 and 0.9%-2.0% on EVAset at PC values of 20% to 60%. In segment overlap measure, these improvements are 1%-2.8% on CB513 at PC values of 20% to 60% and 0.8%-2.2% on EVAset for PC values of 20% to 80%. For PC greater than or equal to 70% but less than 100% the overall accuracy measures of SP1+SP2 becomes close to Homolpro (differing by 0.3%-0.6% on CB513 and 0.3%-0.7% on EVAset) since transferring labels (as in Homolpro) behaves effectively similar to building profile matrices, which are constructed by weighting the templates with respect to similarity scores. A similar behavior is observed for segment overlap measure at PC values greater than or equal to 70 and less than 100% (SP methods differ from Homolpro by 0.1%-0.5% only on CB513 and 0.4%-1.8% on EVAset). When PC threshold is set to 100% the SP1+SP2 performs better than Homolpro by 1% both in overall accuracy and segment overlap measure on CB513, by 1.74% in overall accuracy on EVAset and by 0.29% in segment overlap measure on EVAset. This could be due to the fact that CB513 proteins are older than EVAset and a higher proportion of its targets may have a 90%-100% alignment score with templates in building SP1. However, due to local nature of the HHblits alignments SP1 only may not cover all the residues of the target. The unaligned regions of SP1 may be compensated with the alignments obtained by blastp using proteins with the same PFAM family, which is incorporated into the model through SP2. Comparing the solvent accessibility prediction accuracy of SP1 only and SP1+SP2 with Homolpro, the proposed structural profiling techniques perform better by 0.5%-3.1% on EVAset in overall accuracy for PC values from 20% to 50% and by 1.3%-3.4% for PC values from 60% to 100%. In segment overlap measure, the improvements are 1.2%-5.1% for PC values from 20% to 50% and 4.9%-6.4% for PC values from 60% to 100%. If SP1 only and SP1+SP2 approaches are compared, it can be observed that SP1 only performs slightly better than SP1+SP2 both in CB513 and EVAset for PC threshold less than 60% (i.e. for low to medium similarity levels). At high PC thresholds and especially at PC equal to 100%, SP1+SP2 can perform better than SP1 only in overall accuracy while the segment overlap measure can be higher or lower depending on how well the templates found by PfamScan complements the regions missed by HHblits.

3.5 Detailed accuracy results

Supplementary Sections S6, S7, and S8 include detailed accuracy results obtained for secondary structure and solvent accessibility prediction for the senarios considered in this work. Based on these, the following observations can be made. In secondary structure prediction, the highest improvement is obtained for beta-strands followed by loops and helices when structural profiles are employed. For solvent accessibility prediction, the accuracies of the exposed state improved more than those of the buried state. The amount of improvement is higher as PC threshold is increased from 20% to 60% as compared to the PC region 70%-90%, in which the accuracy values saturate. This is then followed by a large jump in PC equals 100% where even closer templates can be potentially available. Note that such close templates may still match to a sub-region of the target due to local nature of the alignments.

Table 3. Accuracy measures of Homolpro and DSPRED in secondary structure prediction for increasing PC thresholds. A 7-fold cross-validation is performed on CB513.

PC	Homolpro	SP1 only	SP1+SP2	HT1	HT2
20	81.65 (77.30)	82.78 (78.35)	82.59 (78.20)	5.9 (0.0)	5.0 (0.0)
30	81.80 (77.51)	84.21 (80.33)	84.06 (79.74)	13.2 (0.0)	12.3 (0.0)
40	83.29 (79.24)	85.63 (81.42)	85.47 (81.12)	13.2 (0.0)	12.3 (0.0)
50	85.04 (81.22)	86.81 (83.06)	86.65 (82.24)	10.4 (0.0)	9.4 (0.0)
60	86.60 (83.01)	87.70 (84.20)	87.66 (83.69)	6.7 (0.0)	6.5 (0.0)
70	87.43 (83.94)	87.91 (84.25)	88.01 (84.07)	3.0 (0.003)	3.6 (0.0)
80	87.87 (84.62)	87.82 (84.37)	88.31 (84.40)	-0.3 (0.757)	2.8 (0.005)
90	88.29 (84.88)	87.58 (84.36)	88.55 (84.62)	-4.5 (0.0)	1.7 (0.095)
100	91.87 (89.05)	91.67 (88.87)	92.84 (90.22)	-1.5 (0.136)	7.5 (0.0)

Table 4. Accuracy measures of Homolpro and DSPRED in secondary structure prediction for increasing PC thresholds. A 10-fold cross-validation is performed on EVAset.

PC	Homolpro	SP1 only	SP1+SP2	HT1	HT2
20	82.89 (78.24)	83.84 (79.30)	83.75 (79.07)	13.8 (0.0)	12.5 (0.0)
30	82.94 (78.29)	84.70 (80.00)	84.51 (79.82)	25.8 (0.0)	23.0 (0.0)
40	83.58 (78.71)	85.57 (80.91)	85.25 (80.87)	29.8 (0.0)	24.9 (0.0)
50	84.46 (79.48)	86.15 (81.55)	85.71 (81.49)	25.8 (0.0)	19.0 (0.0)
60	85.14 (80.17)	86.68 (82.13)	86.20 (81.98)	23.9 (0.0)	16.4 (0.0)
70	85.66 (80.72)	86.90 (82.48)	86.33 (82.06)	19.5 (0.0)	10.4 (0.0)
80	85.94 (81.04)	86.80 (82.38)	86.33 (81.92)	13.6 (0.0)	6.1 (0.0)
90	86.20 (81.28)	86.69 (82.23)	86.47 (81.72)	7.7 (0.0)	4.2 (0.0)
100	88.88 (84.29)	89.64 (85.57)	90.62 (84.58)	13.3 (0.0)	31.0 (0.0)

4 Conclusion

This work demonstrates that developing more advanced structural profiling methods improves the accuracy of one-dimensional protein structure prediction considerably. Several directions can be considered as future work. First the bit score and e-value score of blastp alignments can be incorporated into the proposed structural profile matrix as multiplicative factors in template weighting equation. Second, a machine learning classifier can be trained that distinguishes whether a target residue has similar structural label with a template residue and scaling the templates using the similarity score prediceted by the classifier. Third, a structural profile matrix database can be constructed and a new structural profile matrix can be derived, which can easily be incorporated into DSPRED. Fourth, target template alignments can be computed using the Smith-Waterman algorithm by incorporating secondary structure,

Table 5. Accuracy measures of Homolpro and DSPRED in solvent accessibility prediction for increasing PC thresholds. A 10-fold cross-validation is performed on EVAset.

PC	Homolpro	SP1 only	SP1+SP2	HT1	HT2
20	79.90 (57.72)	80.57 (59.21)	80.46 (58.96)	9.1 (0.0)	7.6 (0.0)
30	79.90 (57.72)	81.45 (60.57)	81.25 (60.46)	21.2 (0.0)	18.4 (0.0)
40	79.90 (57.72)	82.25 (61.98)	82.02 (61.93)	32.4 (0.0)	29.2 (0.0)
50	79.92 (57.76)	83.02 (62.89)	82.56 (62.83)	43.1 (0.0)	36.6 (0.0)
60	80.13 (58.06)	83.55 (63.73)	83.07 (63.58)	48.0 (0.0)	41.0 (0.0)
70	80.63 (58.92)	83.81 (64.20)	83.24 (63.78)	45.0 (0.0)	36.7 (0.0)
80	80.95 (59.05)	83.82 (64.40)	83.34 (63.94)	40.7 (0.0)	33.7 (0.0)
90	81.23 (59.28)	83.71 (64.92)	83.59 (64.41)	35.3 (0.0)	33.5 (0.0)
100	86.82 (68.12)	87.92 (75.53)	88.93 (74.52)	17.9 (0.0)	34.9 (0.0)

solvent accessibility and torsion angle class predictions into the score update function. Fifth, more advanced alignment techniques that utilize threading can be employed to find better templates and construct more accurate structural profile matrices. Sixth, utilizing templates in secondary structure and solvent accessibility prediction at different similarity levels can be analyzed in terms of its contribution to 3D structure prediction accuracy. Finally, deep learning models that are developed recently for one-dimensional structure prediction such as convolutional and recurrent networks can be incorporated into the models and combined with structural profile information to further improve the prediction accuracy.

Acknowledgements

The experiments reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources) and Feynman Grid Computing Center of CompecTA company.

Funding

This work was supported by 3501 TUBITAK National Young Researchers Career Award [grant number 113E550].

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Aydin, Z., Singh, A., Bilmes, J., and Noble, W. S. (2011). Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. *BMC Bioinformatics*, **12**(1), 154.

Aydin, Z., Baker, D., and Noble, W. S. (2015). Template scoring methods for protein torsion angle prediction. *Communications in Computer and Information Science*, **574**, 206–223.

Cheng, H., Sen, T., Jernigan, R. L., and Kloczkowski, A. (2007). Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: Combining GOR V and Fragment Database Mining (FDM). *Bioinformatics Applications Note*, **23**(19).

Cuff, J. A. and Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**(4), 508–519.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, **44**(D1), D279–D285.

Koh, I. Y. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., Graña, O., Pazos, F., Valencia, A., Sali, A., and Rost, B. (2003). EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Research*, **31**(13), 3311–3315.

Li, D., Li, T., Cong, P., Xiong, W., and Sun, J. (2012). A novel structural position-specific scoring matrix for the prediction of protein secondary structures. *Bioinformatics*, **28**(1), 32–39.

Lin, H.-N., Chang, J.-M., Wu, K.-P., Sung, T.-Y., and Hsu, W.-L. (2005). HYPROSP II A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics*, **21**(15).

Magnan, C. N. and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, **30**(18), 2592–2597.

Mooney, C. and Pollastri, G. (2009). Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins: Structure, Function, and Bioinformatics*, **77**, 181–190.

Pollastri, G., Martin, A. J. M., Mooney, C., and Vullo, A. (2007). Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, **8**(201).

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, **9**(2), 173–175.

Walsh, I., Bau, D., Martin, A. J. M., Mooney, C., Vullo, A., and Pollastri, G. (2009). Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC Structural Biology*, **9**(5).

Wang, G. and Dunbrack, Jr., R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Yang, J. and Zhang, Y. (2016). Protein structure and function prediction using i-tasser. *Current Protocols in Bioinformatics*, **52**, 5.8.1–5.815.

Zhang, W., Liu, S., and Zhou, Y. (2008). SP5: Improving Protein Fold Recognition by Using Torsion Angle Profiles and Profile-Based Gap Penalty Model. *PLoS One*, **3**(6).

Zhou, P., Wen, M., Cong, P., and Li, T. (2017). A predictor of protein secondary structure based on a continuously updated templet library. *Hans Hournal of Computational Biology*, **7**(2), 13–22.