Vol.64: e21210007, 2021
https://doi.org/10.1590/1678-4324-2021210007
ISSN 1678-4324 Online Edition

Article - Engineering, Technology and Techniques

3-State Protein Secondary Structure Prediction based on SCOPe Classes

Sema Atasever[1*]
https://orcid.org/0000-0002-2295-7917

Nuh Azgınoglu[2]
https://orcid.org/0000-0002-4074-7366

Hasan Erbay[3]
https://orcid.org/0000-0002-7555-541X

Zafer Aydın[4]
https://orcid.org/0000-0001-7686-6298

[1]Nevsehir Haci Bektas Veli University, Engineering – Architecture Faculty, Department of Computer Engineering, Nevsehir, Turkey, [2]Kayseri University, Engineering, Architecture and Design Faculty, Department of Computer Engineering, Kayseri, Turkey, [3]University of Turkish Aeronautical Association, Engineering Faculty, Department of Computer Engineering, Ankara, Turkey, [4]Abdullah Gul University, Faculty of Engineering, Department of Computer Engineering, Kayseri, Turkey.

Editor-in-Chief: Alexandre Rasi Aoki
Associate Editor: Raja Soosaimarian Peter Raj

Received: 2021.01.07; Accepted: 2021.06.14.

*Correspondence sema@nevsehir.edu.tr, s.atasever@gmail.com (S.A.).

HIGHLIGHTS

- DSPRED method based on machine learning algorithms to predict 3-state secondary structure elements.
- Comparative results for secondary structure prediction.
- High accuracy of 82.36% based on SCOPe (Structural Classification of Proteins - extended) structural classes.

Abstract: Improving the accuracy of protein secondary structure prediction has been an important task in bioinformatics since it is not only the starting point in obtaining tertiary structure in hierarchical modeling but also enhances sequence analysis and sequence-structure threading to help determine structure and function. Herein we present a model based on DSPRED classifier, a hybrid method composed of dynamic Bayesian networks and a support vector machine to predict 3-state secondary structure information of proteins. We used the SCOPe (Structural Classification of Proteins-extended) database to train and test the model. The results show that DSPRED reached a Q_3 accuracy rate of 82.36% when trained and tested using proteins from all SCOPe classes. We compared our method with the popular PSIPRED on the SCOPe test datasets and found that our method outperformed PSIPRED.

Keywords: Protein secondary structure prediction; SCOPe; Support Vector Machine; Dynamic Bayesian Network.

INTRODUCTION

Proteins are organic molecules that perform certain vital functions in living systems. The structural units of proteins are made up of amino acids, which are connected in sequence such that the carboxyl group of one amino acid forms a peptide bond with the amino group of the next amino acid. To know the structures of proteins is essential to understand the function of proteins. In other words, the structures of proteins reveal crucial information at the molecular level on their functions [1]. The structures of proteins are hierarchically divided into four groups such as primary, secondary, tertiary and quaternary [2]. Each level has its own importance, but the secondary structure is regarded as a bridge between the primary structure and tertiary structure. Thus, protein secondary structure prediction (PSSP) plays a crucial role in the accurate prediction of tertiary structures. As a result, accurate estimation of the secondary structure has been an important research topic for researchers [3]. Protein secondary structure is composed of repeating folding patterns of polypeptide chains. Protein secondary structure is traditionally characterized to be in one of the three general forms such as helix (H), strand (E), and coil (C).

Machine learning algorithms have been used since late 1980s [4]. Recently, their use on PSSP problem is accelerated. Various datasets were used to train and test these models. In this paper, we give emphasis to SCOPe dataset(s) obtained from the SCOPe database, which contains hierarchical classification of proteins based on their structural information. Note that the SCOP dataset version used by the methods developed in the literature can be different from each other and from our version based on the time each study is conducted. Torrisi and coauthors [5] retrained Porter 5 on SCOPe-based dataset. Crooks and coauthors [6] have used a simple Hidden Markov Model (HMM) as an alternative prediction algorithm using the SCOP 1.61 dataset and have found the prediction accuracy, $Q_3 = 65.9 \pm 0.3\%$. Plewczynski and coauthors [7] have found Q_3 score of 73% using the FRAGlib method, which they used as a secondary structure prediction algorithm. Lee and coauthors [8] developed data mining-based RT-RICO model for PSSP. They used SCOP/SCOPe dataset to train and test the model. The model's average Q_3 accuracy was 80.3%. On the other hand, Rashid and coauthors [9] compared the Q_3 accuracies of SCOP classes and revealed the classes with the lowest and highest accuracy rates. They obtained results ranging from 80% to 85% for different SCOP classes in their compact model. Also, in their study, they obtained the best performance from class (a) proteins which have a rich presence of helix residues, and the lowest performance from small proteins with 74% Q_3 accuracy. We have observed that there is a significant similarity between the results of this study and our study. Drozdetskiy and coauthors have used JPred4 which is the latest version of JPred server developed for PSSP. In this study, JNet 2.3.1, a neural network-based predictor, achieved an average Q_3 score of 82% [10]. Yadav and coauthors [11] have identified more than 79% accuracy in the secondary structure of the Human Oxidoreductase family based on SCOP. Martin and coauthors [12] have obtained a 75.5% Q_3 score with multiple sequences using OSS-HMM for secondary structure prediction.

Although experimental techniques yield reliable results on protein structure in laboratory environments, they are labor-intensive, time-consuming, and expensive processes. Thus, various statistical and machine learning computational tools have been developed to predict secondary structure. In this study, we present a model based on DSPRED classifier, a hybrid method composed of dynamic Bayesian networks and a support vector machine to predict 3-state secondary structure elements. We used the SCOPe (Structural Classification of Proteins-extended) database to train and test the model. In recent years, due to improvements in the prediction accuracy of protein secondary structure with the use of hybrid models [13], our article, which is within the scope of machine learning-based hybrid studies, is important in terms of contributing to studies in this field.

MATERIALS AND METHODS

Problem Definition

This study aims to develop a model to determine (i.e. predict) secondary structure states (H: Helix, E: Beta Strands, L: Loops) starting from an amino acid sequence as input. As seen in Figure 1, the first row contains an amino acid sequence which consists of amino acids connected by peptide bonds, and the second row gives the representative secondary structure states (i.e. class labels).

```
Primary   : HLGLFDIRQAIMDYGGLHLQEWCAKGIVNPLFLVRMH
Secondary : LLLLLLHHHHHHHHLLLHHHHHHHHLLLLLLEEEELL
```

Figure 1. Secondary Structure Prediction Problem.

Datasets

SCOPe

In this study, we used SCOPe [14] database version 2.06 available at http://scop.berkeley.edu, which is an extended version of the SCOP [15] database, developed in collaboration with researchers in the Berkeley Lab and UC Berkeley. SCOP is a hierarchically ordered database. Class is the top level of the SCOP hierarchical classification and is based on secondary structure content and organization [16]. It is generated to facilitate the understanding of the historical relationships between proteins and access to available information for protein structures whose structure is known. It incorporates and updates the Astral database. For backward compatibility, data entries in SCOPe are organized in the same hierarchy as SCOP (version 1), see Figure 2. In this study, the experiments were carried out on datasets derived from SCOPe with 40% sequence identity option (i.e. no two proteins have greater than 40% sequence identity). It should be noted that the classification units in SCOP are usually protein domains and a protein domain is a conserved part of the tertiary structure that folds and works independently [15], [17]

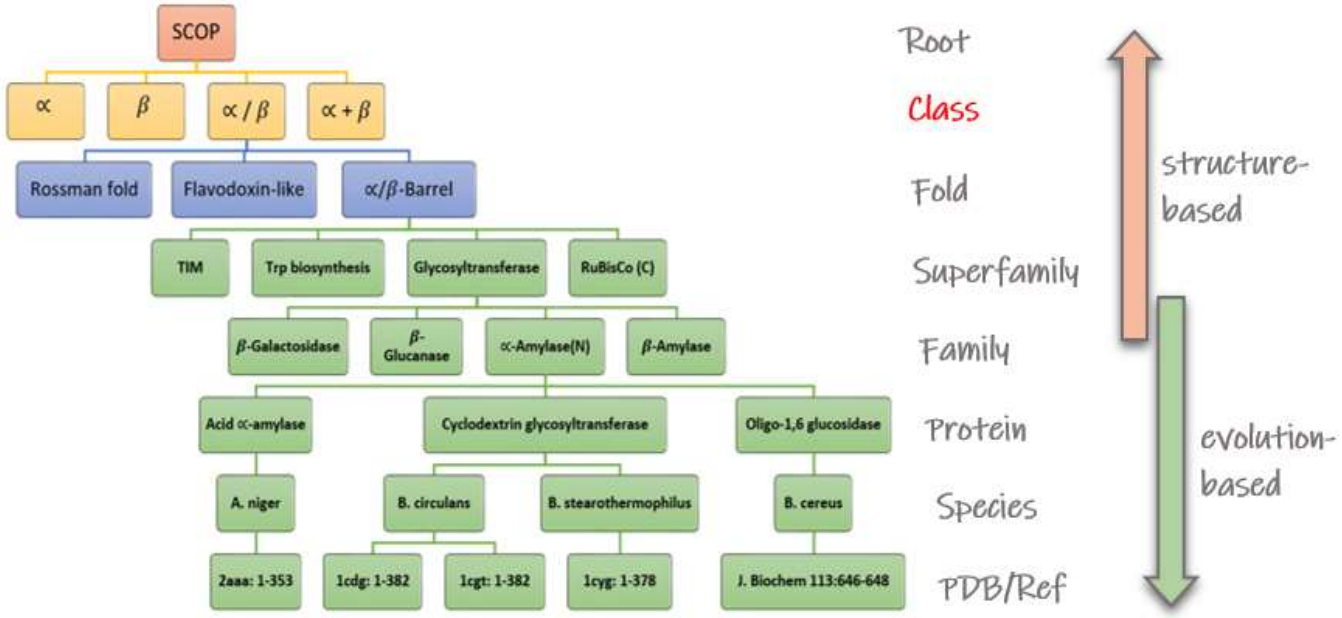

Figure 2. Part of the SCOP (version 1) Hierarchy [18].

The SCOPe database can be accessed from ASTRAL Sequences [19] & Subsets section of the web page's Downloads menu [17]. It is in the form of a dataset, which can be obtained in FASTA [20] format and filtered according to SCCS identifier. Then, protein domain information of the desired classes is obtained. SCCS is the classification string of SCOPe. It is called dot notation and includes SCOPe class, fold, superfamily, and family information.

Non-redundant (NR) database

NCBI maintains the NR nucleotide database for their BLAST search services [21]. NR contains non-identical sequences from various databases such as GenBank and widely used by many researchers for BLAST or Position Specific Iterated-BLAST (PSI-BLAST) search. In NR database, entries with absolutely identical sequences have been merged. NR is available at https://ftp.ncbi.nlm.nih.gov/blast/db/.

PDB99

Protein Data Bank (PDB) [22], [23] digital data archive is widely used for research in structural biology and it also supplies access to 3D structure data for biological macromolecules [24]. PDB structures are available in a plain text file in PDB format. It contains atomic coordinates for biological molecules held in the Protein Data Bank. PDB99 is a database of HMM-profiles generated by Aydin lab and is used in the second step of an HHblits alignment. The proteins in this database was downloaded from the PISCES server by

choosing a threshold value of 99% [25], which is the percentage of sequence identity score, meaning that all protein pairs that have score above this threshold have been eliminated.

Assigning secondary structure labels by DSSP program

In this paper, the true secondary structure labels are assigned by the DSSP [26] program. In order to run this program, first, the PDB files of the proteins in SCOPe should be obtained. For this purpose, the PDB ID and chain information of each protein should be extracted from the stable domain identifier of SCOPe. A seven-character stable domain identifier (sid) contains "d" followed by the 4 character PDB ID, 1-character PDB chain ID and a single character (usually an integer) [17]. For example d4wera1 means chain A of PDB entry 4WER, and d1yjdc1 means chain C of PDB entry 1YJD. First, using sid identifiers (the stable domain identifier) in SCOPe, corresponding PDB ids were parsed from parseable files of database version SCOPe 2.06. After, using PDB entries, we downloaded 3D coordinate information for a particular chain from the PDB database and then we computed the true secondary structure labels by using DSSP starting from the 3D coordinate information.

Reducing 8-states to 3-states

Although DSSP uses an eight-state secondary structure representation for the class label of each amino acid, it is reduced to 3-states: helix (H), beta strands (E) and loop (L) (see Table 1) since prediction methods are usually trained and evaluated for only 3-states [27]. There are different sources where the symbol C for coil is used instead of the letter L for loops [28].

Table 1. Reducing 8-state DSSP labels to 3-states [27]

Description	8-state DSSP labels	3-state DSSP labels
α- helix	H	H
3_{10} – helix	G	
π –helix	I	
β- strand	E	E
β- bridge	B	
β- turn	T	L
Bend	S	
The rest	L or C or ‘ ’	

Comparison with other label assignment programs

In addition to DSSP, there are also other label assignment programs such as STRIDE and DEFINE. Among all three, DSSP [26] and STRIDE [29] are widely used for assigning secondary structure. Note that, the output of these programs may agree at different levels. For instance, in a study made on RS126 benchmark, DSSP and STRIDE agree up to 95%, DSSP and DEFINE agree up to 73% and STRIDE and DEFINE agree at 74% (http://www.compbio.dundee.ac.uk/jpred/references/prot_html/node11.html). In the present work, a comparison of the assignments made by the DSSP and STRIDE programs has also been made (see Supplementary Figure S1 in S3 heading). For this purpose, we randomly selected one protein from each SCOP class and compared the secondary structure assignments made by DSSP and STRIDE. Supporting the previous finding on RS126 dataset, we also found that DSSP and STRIDE assignments are very close to each other with high agreement. Although this is the case, in this paper, DSSP was used as the secondary structure assignment program, as DSSP reference values give better results than STRIDE.

Obtaining train and test sets

Table 2 lists the SCOPe 2.06 structural protein classes used in this study. (https://scop.berkeley.edu/ver=2.06). After the data set belonging to SCOPe was downloaded, six separate FASTA files for each class were created by filtering the FASTA headers in this file according to the SCCS identifier value. These classes include structures with similar secondary structure composition [18]. We then

selected 10% of unique protein domains randomly from each of these six different datasets and produced candidate datasets in FASTA format. For each SCOP class, the remaining 90% of the proteins are included to the corresponding train set. In the next step, proteins in each candidate dataset were aligned with the entire SCOPe dataset using the Blast program, and those with 20% or less identity (remote homologuos) to SCOPe formed the corresponding test dataset, and the remaining proteins (i.e. those that are eliminated from the 10% candidate set during blast alignments) are included to the corresponding the train set (seeTable 2). Finally, we combined the six training sets to form the train set for the all category and combined the six test sets to construct the test set for the all category, which is the union of the six SCOP classes.

Table 2 shows the dataset statistics. In this table, $\mathbf{P_{total}}$ denotes the total number of protein domains belonging to each class, $\mathbf{P_{random}}$ denotes randomly and uniquely selected protein domains, $\mathbf{P_{test}}$ denotes protein domains in test dataset, $\mathbf{P_{train}}$ denotes protein domains in train dataset.

Table 2. Different SCOPe Classes and Randomly Selected Subsets of SCOPe.

Class	Description	P_{total}	P_{random} (10%)	P_{test}	P_{train}
a	All alpha proteins	2423	242	143	2280
b	All beta proteins	2673	267	141	2532
c	Alpha and beta proteins (a/b)	3952	395	157	3795
d	Alpha and beta proteins (a+b)	3329	333	189	3140
e	Multi-domain proteins (alpha and beta)	248	124	49	199
g	Small proteins	667	67	53	613
	Class all (Total)	13292	1428	732	12559

Feature Extraction

PSI-BLAST Alignments

Protein sequence profile search methods have been used as one of the important steps in many bioinformatics studies from past to present. One of the most used and best known of these methods is the BLAST, which contains a set of programs that are used to detect sequence similarities in protein and DNA databases. The PSI-BLAST program is significantly more sensitive than BLAST, but each iteration takes a little more time to run and can be considered as an iterative version of the BLAST [30] algorithm. Many methods developed for PSSP have used the PSSM (Position-Specific Scoring Matrix) [31] which represents the homology information associated with its aligned sequences, which is used as the input feature for prediction [3]. PSI-BLAST can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/.

PSI-BLAST [30] takes an amino acid sequence of a query protein as input and compares it to a protein database. For this purpose, in this study, we have used PSI-BLAST to obtain profile matrices (i.e. position specific scoring matrix called PSSM) for each protein sequence in the SCOPe dataset. First, we have executed PSI-BLAST version 2.3.0+ with the following parameters: number of iterations=3, e-value threshold=10, inclusion threshold=0.001 and the number of threads=16. Using the PSI-BLAST program, we have searched the non-redundant sequence database NR from the National Center for Biotechnology Information (NCBI). PSI-BLAST produces a text-based alignment file and an Nx20 PSSM for each protein domain in the SCOPe dataset. The scores in each PSSM are then normalized to interval [0,1] by applying a sigmoidal transformation (i.e. by passing the score through a sigmoid function and retrieving the output score).

HHblits Alignments

HH-suite, which is a free, open source software suit, is often used for highly sensitive sequence searching and protein structure prediction. HH-suite contains HHsearch and HHblits, which is an accelerated version of HHsearch. HHBlits converts the query sequence to an HMM and iteratively searches through standard HHblits databases such as UniProt20, NR20, and uniclust30. Compared to PSI-BLAST, HHblits is faster [32], [33]. An HMM-profile can model protein sequences, which can be used as input features for the purpose of protein structure prediction [33], [34]. The latest version of HH-suite is available at https://github.com/soedinglab/hh-suite.

We have used HHblits (from HHsuite 2.0.16) to construct the alignments. We have aligned the SCOPe test proteins to the NR20 database (which is a subset of the NR database filtered using a 20% sequence identity threshold) and have generated HMM-profile models (first step), which are then aligned to the HMM-profile models in the PDB99 database (second step).

Generating Structural Profile Matrices

Structural profile matrices (SPMs) with .struct extension were obtained for each protein domain in the dataset using files with .hhr extension obtained using the HH-suite program. Figure 3 shows a section of an example SPM. Each row in an SPM contains three scores (values from 0 to 1) that represent the tendency of each amino acid of the target to be in one of the 3-state secondary structure classes. The sum of the scores in each row is 1. The size of an SPM generated for secondary structure prediction is N x 3, where U is the number of amino acids in the target protein and each column represents one of the three secondary structure states: H = helix, E = beta-strand, L = loop.

```
0.750394 0.124803 0.124803
0.135899 0.018704 0.845398
0.982206 0.008897 0.008897
0.984268 0.007866 0.007866
0.985902 0.007049 0.007049
0.986025 0.006987 0.006987
0.986025 0.006987 0.006987
0.986147 0.006927 0.006927
0.986147 0.006927 0.006927
0.986147 0.006927 0.006927
0.986790 0.006605 0.006605
0.987545 0.006227 0.006227
```

Figure 3. A subsection of an SPM.

Prediction Method

In this study, the PSSP was made with the DSPRED [35], [36] method. The name DSPRED is short for DBN SVM predictor. The letter D stands for the Dynamic Bayesian Networks (DBN), and the letter S stands for the Support Vector Machine (SVM). DSPRED method is available at: https://github.com/yusufzaferaydin/dspred and the web server version is available at http://psp.agu.edu.tr.

DSPRED is a two-stage hybrid method, the first stage of which contains DBN classifiers and the second stage contains an SVM. In the DBN stage, a separate DBN model is trained for PSI-BLAST and HHMAKE PSSMs, see Figure 4. Using the previously generated profile matrices for each class in the SCOPe dataset, the Dynamic Bayesian Networks were trained and predicted probability score distributions 1-2 which the prediction results from this stage are combined with an SPM obtained from the second stage of the HHBlits method to compute distribution 3. In the next step, the PSSM features and all three distributions are forwarded as input to an SVM. The datasets were divided into training and testing sets as explained in Section Feature Extraction and summarized in Table 2. This includes assigning (i.e. splitting) proteins to train and test sets.

After this assignment is made, input feature sets (2D arrays with rows representing amino acids and columns denoting features) and output labels of the amino acids are formed for train and test sets of the SVM classifier. For this purpose, a two-fold cross-validation is performed on each train set to compute the probability distributions 1-3 of the DBN step (Figure 4), which are used to form feature set of the SVM model for train set proteins when combined with the PSSM features (Figure 4). In the next step, DBN models are trained on the full train set and distributions 1-3 are obtained for the corresponding test set, which are combined with the PSSM features to form the feature set of the SVM model for test set proteins. A symmetric window of size 11 is taken around each amino acid and the corresponding feature vectors coming from PSSMs and distributions 1-3 are concatenated to form the feature set of a given amino acid. Once the input features and output labels for the train and test sets of the SVM model are formed, model training is performed on each train set and predictions are computed on the corresponding test set of the SVM. A separate train set/test set pair is formed for each SCOP class tabulated in Table 2.

The DBN models are implemented using the graphical models toolkit (GMTK) [37] and the SVM classifier is implemented using the libSVM [38] (version 3.21) and ThunderSVM software [39], which is the GPU accelerated version of libSVM. To be more specific, for the all category in Table 2, SVM model is trained using ThunderSVM (since the all class category contains the largest train set) and for the remaining SCOP classes libSVM is used. The reason for this choice is because although SVM is a powerful classifier, it is not suitable for large datasets due to quadratic optimization involved in model training. Since the training dataset for the all category in Table 2 is large, ThunderSVM is used for this dataset (and also for the corresponding test set) on a high performance computing (HPC) system. ThunderSVM can use the high-performance of Graphics Processing Units (GPUs) or multi-core CPUs [39]. It performs acceleration by parallelizing the kernel computation step of SVM.

As a widely-used choice for small to moderate feature set sizes [40], we used the RBF kernel in our SVM model with hyper-parameters set to C=1.0 and α = 0.00781.

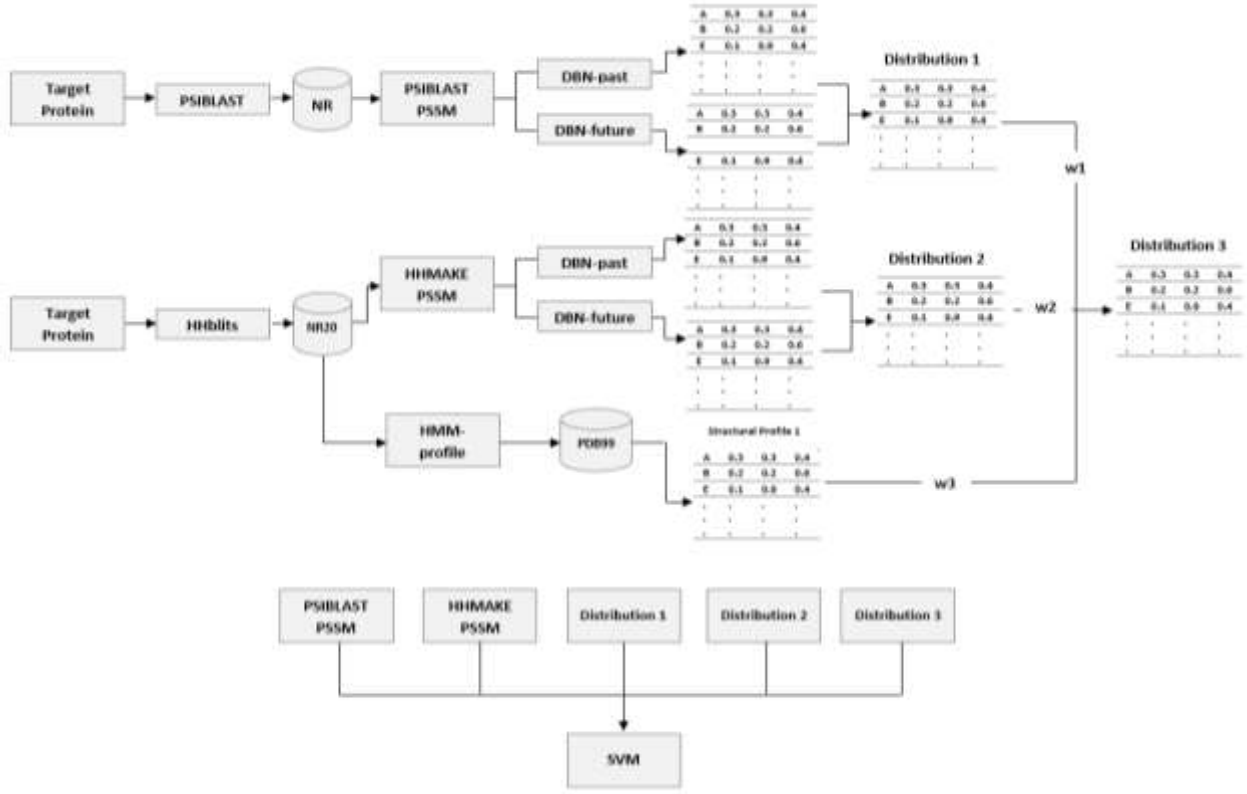

Figure 4. Overall pipeline of our DSPRED method.

System Architecture

The CPU-based calculations (i.e. DSPRED with libSVM and PSIPRED) are run on a Centos Enterprise Linux 7.3 OS, with an 2x Intel Xeon i5-2690 (16 cores 32 threads in total) CPU and 256 GB 1600MHz ECC RAM as well as on Ubuntu 16.04.2 LTS (Xenial Xerus) OS, with an 32 CPUs, 2 sockets, 8 cores per socket, 2 threads per core, 2 nodes, CPU model Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz, and 64 GBs of RAM. The GPU-based calculations (i.e. DSPRED with ThunderSVM) are run on NVIDIA DGX-1 Tesla V100 with

128 GB GPU RAM, Dual 20-core Intel® Xeon® E5-2698 v4 2.2 GHz CPU, 40960 CUDA cores, and 5120 Tensor cores.

RESULTS

Throughout this section, first, the evaluation metrics used in this study are described. Second, the experimental results obtained are presented in graphical form. Detailed version of all results including the confusion matrices can be found in Supplementary Section S1. Third, the current study is compared with the existing state-of-the-art studies that used the SCOPe database.

Evaluation Criteria

As evaluation metrics, in addition to overall accuracy of Q_3 [41] and Segment Overlap Measure (SOV) [42], class-specific recall, precision and Matthew's Correlation Coefficient (MCC) [43] measures as well as the confusion matrix are also used to assess the performance of the models. The overall accuracy (i.e. Q_3) refers to the ratio of accurate predictions to all predictions (i.e. the sum of all amino acids in test set). The recall value is calculated for each class type as the number of the structural labels estimated correctly divided by the total number of actual structural labels belonging to that class type. For example, the recall for helix class is the number of correctly predicted helices divided by the number of actual helices. Precision is also calculated separately for each class type as the number of structural labels estimated correctly divided by the total number of predictions for that class type. SOV is a segment-based evaluation criterion used especially for protein secondary structure prediction. Although the ranking of the prediction methods based on SOV scores is similar to Q_3, SOV is a more sensitive and realistic assessment method in terms of prediction quality [42], [44] and typically takes lower values as compared to the residue-based Q_3 measure due to the fact that the accuracy is computed at segment level, which is more stringent than a residue-based evaluation. The Matthews correlation coefficient is a measure of quality of binary classifications. It is a correlation coefficient between the observed and predicted binary classifications [45]. In this work, MCC is computed for each class-type separately in a one-versus-rest setting. Confusion matrix contains number of correct predictions and errors for different secondary structure labels.

Prediction accuracy of DSPRED for SCOPe classes and comparison to PSIPRED

In this section, the prediction performance of DSPRED is compared to PSIPRED [46], which is one of the popular secondary structure prediction methods. DSPRED is trained and tested separately for each of the SCOPe classes given in Table 2. For example, for SCOPe class "a", the training and test sets of DSPRED included proteins belonging to this SCOPe class only. Secondary structure predictions for the test sets of different SCOPe class categories in Table 2 are computed using the stand-alone version of PSIPRED version 4.02 without performing any additional model training (since the trainable version of PSIPRED is not available for download). Table 3 includes the overall Q_3 accuracy and SOV metrics of DSPRED and PSIPRED for various SCOPe classes. Other metrics such as precision, recall, MCC, and confusion matrices that contain various types of errors between different secondary structure labels are included in Supplementary Section S2. Based on these results, even if the stand-alone version of PSIPRED is used (allowing overlaps between the test sets and the train set of PSIPRED), our method DSPRED performs better than PSIPRED except for the SOV measure for the all class. Although both methods are two-stage approaches, the reason for having a better performance using DSPRED can be mainly associated with the use of HHMAKE PSSM and structural profile matrices derived by HHblits as input features. The overall Q_3 accuracy of DSPRED reached to 82.36% and the SOV accuracy to 72.55%. The highest Q_3 accuracy of 87.30% is obtained for Class a and the lowest Q_3 accuracy of 78.10% for Class g. Figure 5 shows the distribution of the protein-level accuracies for DSPRED for the SCOPe all class.

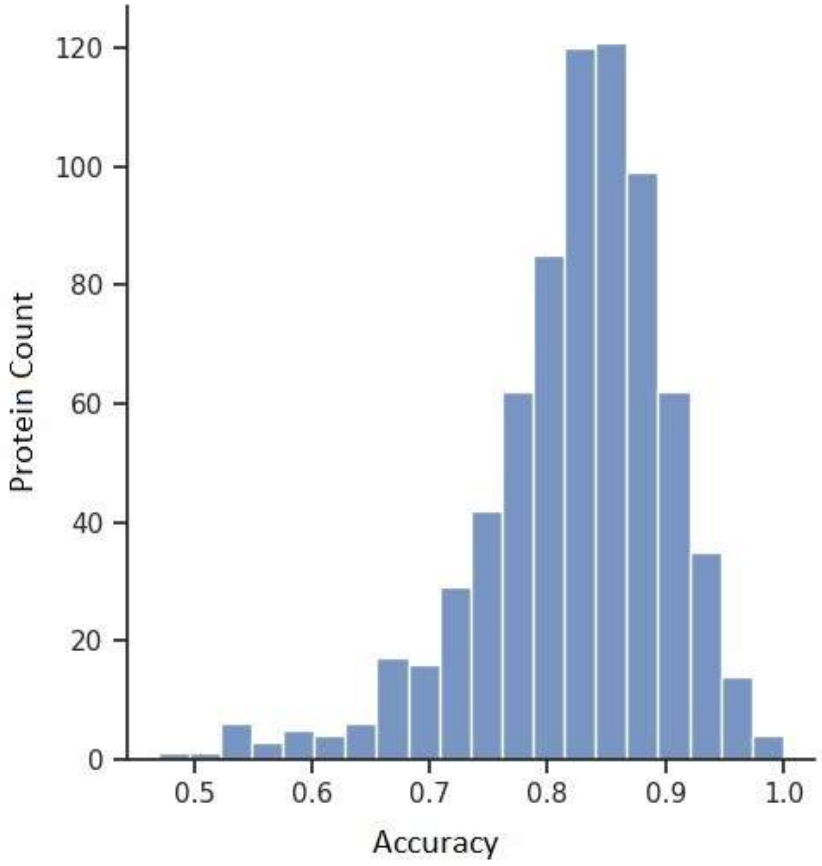

Figure 5. Distribution of the protein-level accuracies for DSPRED for the SCOPe all class

Table 3. A comparison between the overall Q_3 accuracy and SOV metrics of DSPRED and PSIPRED.

		DSPRED (This work)		PSIPRED	
Class	**Description**	**Q_3**	**SOV**	**Q_3**	**SOV**
a	All alpha proteins	87.3	82.2	85.2	80.22
b	All beta proteins	80.39	73.95	78.02	72.85
c	Alpha and beta proteins (a/b)	85.08	81.72	81.92	78.74
d	Alpha and beta proteins (a+b)	83.34	78.89	80.1	76.46
e	Multi-domain proteins (alpha and beta)	80.5	74.01	78.9	72.78
g	Small proteins	78.1	66.63	71.92	62.59
Class all (Total)		82.36	72.55	80.72	76.27

DISCUSSIONS

Estimating SCOP class for PSSP

In order to use the versions of DSPRED trained using proteins that belong to specific SCOP classes only, it is necessary to know the SCOP class of the target protein, which may not be available in general. In our experiments with the individual SCOP classes (i.e. categories a to g) we assumed that this information is available. If the structural class information is not available then we can perform prediction by DSPRED trained using a large dataset that includes proteins from all SCOP classes. This is performed in the seventh train-test experiment for the all category. As an alternative, it may be possible to estimate the SCOP class of the target protein using various techniques and use the model that is trained with proteins belonging to that particular SCOP class only to make predictions. One technique to estimate the SCOP class could be to perform alignment with HMM-profiles of SCOP proteins using HHblits and taking a majority voting of the SCOP classes of the hit proteins in top scoring alignments. There are also other techniques in the literature to estimate the SCOP class of a protein, which can be used to select the DSPRED version trained using the appropriate SCOP class only. For instance, it can be possible to use chemical shift information from nuclear magnetic resonance experiments (when available) to predict structural class as well as secondary structure of proteins. We leave exploring these directions as a future work, which may help us to develop our prediction method further.

Structural variability

In this paper, we framed secondary structure prediction as a single-label classification problem, which assumes that there can be one true label only for each amino acid (i.e. one true fold for a protein sequence). These labels were derived using the DSSP program. However, it is known in the literature that protein structures can be dynamic and can assume multiple folding states. Furthermore, there is no single definition for assigning secondary structure labels starting from a single 3D coordinate file and as a result multiple secondary structure assignment programs have been developed such as DSSP, STRIDE, and DEFINE. These factors may cause the protein to have multiple secondary structure state sequences. In order to capture such variations, a better approach could be to develop and train a multi-label prediction model, which allows one to assign multiple secondary structure labels to each amino acid. For this purpose, the true labels of secondary structure can be derived using different assignment programs such as DSSP, STRIDE, and DEFINE. Although a multi-label version of DSPRED may not be developed in a straightforward manner, alternative approaches can be followed such as training separate DSPRED versions for each of the label assignments made by DSSP, STRIDE and DEFINE and computing separate predictions for these assignments. Then the accuracy can be computed in a multi-label setting. Alternatively, deep neural networks can be trained in a multi-label and a multi-class setting allowing each amino acid to have multiple class labels.

In the present study, the HMM-profiles of target proteins are aligned with HMM-profiles of proteins in PDB99 database using HHblits. The PDB99 database contains more than thirty thousand proteins, which can be considered as large but is derived as a non-redundant version of the PDB database by eliminating protein pairs with more than 99% sequence identity. As a result, it does not include all proteins in PDB, where a limited number of structures may be kept representing each protein family. This may be a limiting factor for capturing the structural variations of the target proteins (e.g. those caused by mutations). It is known that even single mutations may have large impacts on the folding of proteins. Therefore in order to increase the variability of the structure database used for HHblits, PDB99 can be further extended to include more PDB proteins. This may allow us to have more relevant hits per target protein and results in better PSSM and better structural profile matrix computation.

CONCLUSIONS

In this paper, we have presented a hybrid method for PSSP based on machine learning methods trained and tested by using SCOPe datasets. We showed that our method outperformed PSIPRED on SCOPe test datasets. In addition to the extensions mentioned in discussion section for future studies, combining different structural profile matrices with deep learning methods can also be considered in the future and better results can be obtained for further improving the performance.

Funding: This work was supported by 3501 TUBITAK National Young Researches Career Award [grant number 113E550].
Acknowledgments: The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).
Conflicts of Interest: The authors declare no potential conflict of interest.

REFERENCES

1. Rigden DJ. From protein structure to function with bioinformatics. Rigden DJ, editor. Berlin: Springer; 2009.
2. Linderstrøm-Lang KU. Lane medical lectures: proteins and enzymes. Stanford University Press; 1952.
3. Ma Y, Liu Y, Cheng J. Protein secondary structure prediction based on data partition and semi-random subspace method. Scientific reports. 2018 Jun 29;8(1):1-0.
4. Juan SH, Chen TR, Lo WC. A simple strategy to enhance the speed of protein secondary structure prediction without sacrificing accuracy. PloS one. 2020 Jun 30;15(6):e0235153.
5. Torrisi M, Kaleel M, Pollastri G. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. Scientific reports. 2019 Aug 26;9(1):1-2.
6. Crooks GE, Brenner SE. Protein secondary structure: entropy, correlations and prediction. Bioinformatics. 2004 Jul 1;20(10):1603-11.
7. Plewczynski D, Rychlewski L, Ye Y, Jaroszewski L, Godzik A. Integrated web service for improving alignment quality based on segments comparison. BMC bioinformatics. 2004 Dec;5(1):1-7.
8. Lee L, Leopold JL, Frank RL, Maglia AM. Protein secondary structure prediction using rule induction from coverings. In 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2009 Mar 30 (pp. 79-86). IEEE.
9. Rashid S, Saraswathi S, Kloczkowski A, Sundaram S, Kolinski A. Protein secondary structure prediction using a small training set (compact model) combined with a Complex-valued neural network approach. BMC bioinformatics. 2016 Dec;17(1):1-8.
10. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic acids research. 2015 Jul 1;43(W1):W389-94.
11. Yadav BS, Pokhariyal M, Ratta B, Rai G, Saxena M. Predicting Secondary Structure of Oxidoreductase Protein Family Using Bayesian Regularization Feed-forward Backpropagation ANN Technique. J Proteomics Bioinform. 2010;3:179-82.
12. Martin J, Gibrat JF, Rodolphe F. Analysis of an optimal hidden Markov model for secondary structure prediction. BMC structural biology. 2006 Dec;6(1):1-20.
13. Jiang Q, Jin X, Lee SJ, Yao S. Protein secondary structure prediction: A survey of the state of the art. Journal of Molecular Graphics and Modelling. 2017 Sep 1;76:379-402.
14. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014 Jan 1;42(D1):D304-9.
15. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 1995 Apr 7;247(4):536-40.
16. Getz G, Vendruscolo M, Sachs D, Domany E. Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. Proteins: Structure, Function, and Bioinformatics. 2002 Mar 1;46(4):405-15.
17. SCOPe. [Internet]. [cited: 2020 Sep 19]. Available from: https://scop.berkeley.edu
18. SCOP 1.75 help. [Internet]. [cited: 2020 Dec 3]. Available from: https://scop.berkeley.edu/help/ver=1.75
19. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res. 2000 Jan 1;28(1):254-6.
20. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences. 1988 Apr 1;85(8):2444-8.
21. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2006 Jan 1;34(suppl_1):D173-80.
22. Bernstein FC, Koetzle TF, Williams GJ, Meyer Jr EF, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.. 1977 May 25;112(3):535-42.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42.
24. The RCSB Protein Data Bank. [Internet]. [cited: 2020 Sep 2]. Available from: http://www.rcsb.org/
25. Wang G, Dunbrack Jr RL. PISCES: a protein sequence culling server. Bioinformatics. 2003 Aug 12;19(12):1589-91.

26. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules. 1983 Dec;22(12):2577-637.
27. Uzut, ÖG. Optimizing Classifiers For Protein Secondary Structure Prediction [dissertation]. Abdullah Gul University; 2017.
28. Hofmann DW. Data mining in organic crystallography. Data mining in crystallography. 2009:89-134.
29. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function, and Bioinformatics. 1995 Dec;23(4):566-79.
30. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.
31. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 1999 Sep 17;292(2):195-202.
32. Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2005 Apr 1;21(7):951-60.
33. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. methods. 2012 Feb;9(2):173-5.
34. Bystroff C, Krogh A. Hidden Markov Models for prediction of protein features. Protein Structure Prediction. 2008:173-98.
35. Aydin Z, Singh A, Bilmes J, Noble WS. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. BMC bioinformatics. 2011 Dec;12(1):1-21.
36. Atasever S, Aydın Z, Erbay H, Sabzekar M. Sample Reduction Strategies for Protein Secondary Structure Prediction. Appl. Sci. 2019 Jan;9(20):4429.
37. Bilmes J, Zweig G. The graphical models toolkit: An open source software system for speech and time-series processing. In2002 IEEE International Conference on Acoustics, Speech, and Signal Processing 2002 May 13 (Vol. 4, pp. IV-3916). IEEE.
38. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011 May 6;2(3):1-27.
39. Wen Z, Shi J, Li Q, He B, Chen J. ThunderSVM: A fast SVM library on GPUs and CPUs. J. Mach Learn Res. 2018 Jan 1;19(1):797-801.
40. Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Taipei; 2003.
41. Clementi C, Garcıa AE, Onuchic JN. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. J. Mol. Biol. 2003 Feb 21;326(3):933-54.
42. Zemla A, Venclovas Č, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins: Structure, Function, and Bioinformatics. 1999 Feb 1;34(2):220-3.
43. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1975 Oct 20;405(2):442-51.
44. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 2001 Apr 27;308(2):397-407.
45. Banerjee S. Prediction of crystal packing and biological protein-protein interactions with Linear Dimensionality Reduction-SVD.Banerjee S. Prediction of crystal packing and biological protein-protein interactions with Linear Dimensionality Reduction-SVD.
46. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000 Apr 1;16(4):404-5.