Accepted Manuscript

Journal of Bioinformatics and Computational Biology

Article Title:	Dimensionality Reduction for Protein Secondary Structure and Solvent Accesibility Prediction
Author(s):	Zafer Aydin, Oguz Kaynar, Yasin Gormez
DOI:	10.1142/S0219720018500208
Received:	01 August 2018
Accepted:	01 August 2018
To be cited as:	Zafer Aydin, Oguz Kaynar, Yasin Gormez, Dimensionality Reduction for Protein Secondary Structure and Solvent Accesibility Prediction, *Journal of Bioinformatics and Computational Biology*, doi: 10.1142/S0219720018500208
Link to final version:	https://doi.org/10.1142/S0219720018500208

World Scientific
www.worldscientific.com

Journal of Bioinformatics and Computational Biology

DIMENSIONALITY REDUCTION FOR PROTEIN SECONDARY STRUCTURE AND SOLVENT ACCESIBILITY PREDICTION

Zafer Aydın

Department of Computer Engineering, Abdullah Gul University
Kayseri, 38080, Turkey
zafer.aydin@agu.edu.tr

Oğuz Kaynar

Department of Management Information Systems, Cumhuriyet University
Sivas, 58000, Turkey
okaynar@cumhuriyet.edu.tr

Yasin Görmez*

Department of Management Information Systems, Cumhuriyet University
Sivas, 58000, Turkey
yasingormez@cumhuriyet.edu.tr

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Secondary structure and solvent accessibility prediction provide valuable information for estimating the three dimensional structure of a protein. As new feature extraction methods are developed the dimensionality of the input feature space increases steadily. Reducing the number of dimensions provide several advantages such as faster model training, faster prediction and noise elimination. In this work, several dimensionality reduction techniques have been employed including various feature selection methods, autoencoders and PCA for protein secondary structure and solvent accessibility prediction. The reduced feature set is used to train a support vector machine at the second stage of a hybrid classifier. Cross-validation experiments on two difficult benchmarks demonstrate that the dimension of the input space can be reduced substantially while maintaining the prediction accuracy. This will enable the incorporation of additional informative features derived for predicting the structural properties of proteins without reducing the accuracy due to overfitting.

Keywords: Secondary Structure Prediction; Solvent Accessibility Prediction; Feature Selection; Dimension Reduction; Autoencoder.

1. Introduction

Protein structure prediction is one of the most important and challenging problems of computational biology. Accurate prediction of three-dimensional (3D) structure provides important information about the function of a protein. Furthermore it is an effective

* Corresponding author.

alternative to experimental methods which are costly and time consuming. Structure prediction is also important for drug and enzyme design.

Several one-dimensional (1D) structural characteristics such as secondary structure, profile matrices, torsion angles, and solvent accessibility are used as features to predict the 3D structure of a protein. Inferring these with minimum error is important for accurate 3D structure prediction.

To date many machine learning algorithms have been developed for predicting 1D properties of proteins. Among those support vector machines (SVM) is widely used. Hua and Sun applied an SVM on RS126 and CB513 datasets and obtained a 73.5% Q3 score for protein secondary structure prediction.[1] Aydin et al. used SVM and Dynamic Bayesian networks on CB513 and achieved 80.3% Q3 score.[2] Huang and Chen used support vector machines on a dataset that is generated using PSSM values and four physicochemical features (net charges, conformation parameters, side chain mass, and hydrophobic), and obtained a Q3 accuracy of 79.52%.[3] Wang et al. optimized the parameters of support vector machines by grid search and genetic algorithm. In this study, the model trained by using grid search had 76.08% Q3 score and the model learned by using genetic algorithm produced a 76.11% as the Q3 score.[4] Solvent accessibility is another problem that is studied for protein structure prediction. Pugalenthi et al. developed a 5-random-forest model with five different thresholds for buried and exposed states and reached a 77.57% prediction accuracy.[5] Joo et al. used the nearest neighbor method to predict 2 and 3 states of solvent accessibility and obtained a 78.38% accuracy for 2-state and 65.1% accuracy for 3-state prediction.[6] Adamczak et al. developed a regression model to predict continuous valued solvent accessibility and a neural network model for two class classification. They obtained 15.3-15.8% mean absolute errors for the regression model and 77% prediction accuracy for the neural network model.[7] Faraggi et al. proposed a guide-learning method for residue solvent accessibility and they reduced the mean absolute error by 2-4%.[8]

Efforts in predicting structural properties of proteins are not limited to developing advanced classification methods. The first secondary structure prediction algorithms were based on the tendency of each amino acid to form helices or leaves, and rules to estimate the formation of secondary structural elements. These methods had 60% accuracy in predicting which of the three states (helix, strands, loop) an amino acid residue would adopt. Subsequently, a significant increase in accuracy was achieved by using multiple sequence alignments and the success rate reached 80-82%.[2, 9] When structural profiles were used to summarize information contained in templates, the accuracy further increased to 84-85%.[10, 11] Though most of the studies used PSI-BLAST profiles as input features it has been shown that incorporating HHblits profiles and structural profiles improves the accuracy of classification.[2, 12] While extracting informative features provides a promise for improving the accuracy of structure prediction, adding new features to the feature set will eventually increase the dimensionality of the input space, which could be harmful for the accuracy and performance of a classifier. Having too many features may increase the training time and can cause overfitting, which reduces the

accuracy on unseen data. Furthermore it can distort training due to noisy features. On the other hand a few features may not be sufficient for training, which is also known as underfitting. Hence, proper and sufficient number of features should be employed in machine learning models. To solve the aforementioned problems, dimensionality reduction techniques such as feature selection and projection methods can be used.[13]

Despite numerous methods developed for predicting 1D properties of proteins, there is limited work on reducing the number of dimensions in order to incorporate new feature descriptors. Li et al. applied principal component analysis on 3 benchmark datasets and obtained 86.7% Q3 accuracy by support vector machines.[14] Adamczak used t-statistics and information gain for feature selection and principal component analysis for dimension reduction. He trained a neural network with reduced data set and achieved a 79.1% Q3 accuracy.[15]

In this work, principal component analysis (PCA),[16] autoencoder (AE),[17] ranker chi-square (X^2),[18] ranker information gain (IG),[19] ranker gain ratio (GR),[20] minimum redundancy maximum relevance (MRMR),[21] correlation-based genetic feature selection (CFS-ge),[22,23] correlation-based greedy feature selection (CFS-gr)[22,24] and correlation-based best first feature selection (CFS-bf)[22,25] are used as dimension reduction techniques for protein secondary structure (PSSP) and solvent accessibility (SA) prediction. A support vector machine from a two-stage classifier is employed to predict the secondary structure and solvent accessibility class of amino acids. A rich feature set representation is used that includes PSI-BLAST PSSM profiles, HHMAKE PSSM profiles and structural profiles. To the best of our knowledge there is no work in the literature that compares different dimension reduction methods in terms of accuracy and speed for protein secondary structure and solvent accessibility prediction starting from the aforementioned feature set.

2. Definitions and methods

2.1. *Secondary structure prediction*

The aim of this problem is to assign a secondary structure class (helix, beta strand, or loop) to each amino acid of a protein (Fig. 1). To estimate secondary structure typically supervised learning approaches are used, in which a model is trained by using proteins that have known secondary structure.

M	S	N	T	T	W	G	L	Q	R	D	I	T	P	R	L
L	L	E	H	H	E	H	H	E	L	E	H	E	L	H	E

Fig. 1 Three state protein secondary structure prediction. The first line is the amino acid sequence. The second line is the 1D secondary structure representation.

2.2. *Solvent accessibility prediction*

Solvent accessibility prediction can be used to determine which amino acids are on the outer surface of the protein and which are in the inner region of the protein. This can

provide several constrains for 3D structure prediction. The aim of solvent accessibility prediction may be either to calculate the real-valued solvent accessibility score or to predict the accessibility class. In the literature the second problem has been studied more than estimating real-valued scores. Because, the accessible surface area can have different values for different amino acids, it is converted to relative solvent accessibility as a result of a standardization process. For this purpose, the accessible surface area of each amino acid calculated by the DSSP program[26] is divided by the maximum accessibility of that amino acid. Relative solvent accessibility information is more useful for predicting 3D structure than standard solvent accessibility. Relative solvent accessibility values are then transformed into discrete accessibility classes. Several accessibility classes have been proposed in the literature. Commonly 2, 3 and 4 class representations are used. For example, in the definition of 2-state of accessibility, each amino acid in the training set is assigned to one of the exposed or buried classes. For this purpose, continuous–valued accessibility scores are compared against a threshold, which can take values such as 0%, 5%, 10%, 25%, and 50%. In this paper, 25% is used as the threshold value. Once the label information is obtained, a prediction model can be trained to estimate the accessibility class labels for proteins with unknown structure.

2.3. *Feature extraction for one dimensional protein structure prediction*

Proteins with similar amino acid sequences typically have similar structure. When the amino acid sequence is different, there is usually no structural similarity. However, there are proteins, which have similar structures but different amino acid sequences. To capture the information contained in this type of variation, statistical profiling techniques are employed. Among those the profile matrix such as position specific scoring matrix (PSSM) is mainly a statistical score table obtained by aligning proteins in the same family. It contains a likelihood score for observing the 20 amino acids in each position of the query protein. In this study, we use PSI-BLAST PSSM,[27] HHMAKE PSSM,[28] and structural profile matrices as input features to predict secondary structure and solvent accessibility of proteins.[12] The PSI-BLAST PSSMs are obtained by running the PSI-BLAST method against the NR database, HHMAKE PSSMs are obtained by aligning the targets against the NR20 database using the first step of the HHblits method (which also produces HMM-profile for target) and structural profiles are obtained by aligning the HMM-profiles of the target against the PDB[29] database using the second step of the HHblits method and by normalizing the weighted frequencies of the structural label information of PDB proteins. Details of weighted frequency computation can be found in the paper by Aydin et al.[12] In the present study and in Aydin et al.[2], only distant templates are used to construct structural profiles matrices. This is achieved by removing templates for which the percentage of sequence identity score with respect to query is greater than 20%. Once the profile matrices are obtained they are scaled by sigmoid transformation as in Aydin et al. to map the features to range $[0,1]^2$ and sent as input to a classifier. Fig. 2 summarize the steps of feature extraction and DSPRED method.

2.4. *Classification methods*

2.4.1. *DSPRED method*

The DSPRED is a two-stage classifier that includes Dynamic Bayesian Networks (DBN)[30] and a support vector machine classifier.[2] A separate DBN is trained for PSIBLAST PSSM and HHMAKE PSSM features. DBN models produce marginal a posteriori probability distributions of structural class labels given features for each amino acid of the target. These are denoted as Distribution 1 and Distribution 2, respectively. In the next step, Distribution 3 is computed as the average of Distribution 1, Distribution 2, and structural profile matrix. Note that due to local nature of HHblits alignments, some columns of the structural profile matrix may not contain any hits from PDB proteins. For such amino acids of the target, Distribution 3 is computed as the average of Distribution 1 and 2 only. In the second stage of DSPRED, the profile matrices (PSI-BLAST and HHMAKE) used for DBN are combined with Distributions 1, 2, and 3 and sent to a support vector machine[31] as input. For this purpose, a sliding and symmetric window around each amino acid is selected and the columns of the profile matrices, Distributions 1, 2, and 3 that correspond to these windowed positions are used as input features. Finally, the support vector machine predicts the secondary structure or solvent accessibility class of the amino acid at the center of the window. In this paper, we used the one-versus-one SVM technique[32] for three-class protein secondary structure prediction (fig. 2). There is no feature selection or dimension reduction in the original DSPRED method, but in this study these methods are applied before SVM.

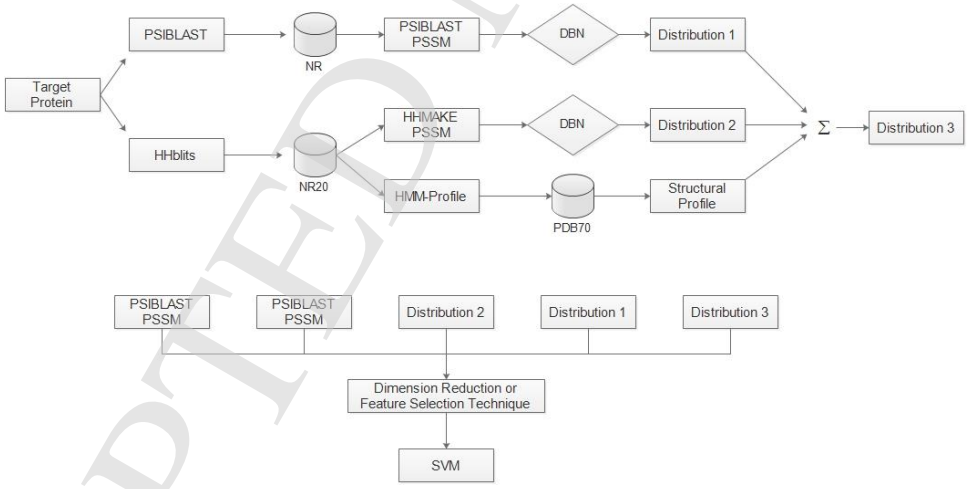

Fig. 2. Steps of feature extraction and DSPRED method for 1D protein structure prediction

2.5. *Feature selection techniques*

Feature selection methods can be categorized into two groups: filter algorithms and wrapper algorithms. In the filter approach, a score is computed for each feature, which enables the features to be ranked. Then those features that have a score smaller than a threshold are eliminated. Filter based methods can also be accompanied with a search algorithm that is used to find the optimum feature subset. In this paper, Correlation-based feature selection (CFS) is employed as the filtering approach and Best First[24], Genetic[22] and Greedy[23] are employed as the search methods. In the wrapper approach, features are selected by using a ranking method, which can also be combined with a search method, and a classifier that optimizes the prediction accuracy. For the wrapper methods of this work, features are ranked using Chi-square, Information Gain, or Gain Ratio and then optimum feature subset is found using SVM as the classification method. Finally in minimum redundancy maximum relevance algorithm firstly features are ranked by using MRMR metric[21] then the best features are selected by forward feature selection using correlation based feature selection algorithm.

2.6. *Projection techniques*

In projection-based dimensionality reduction, features are mapped to a new space in which the number of dimensions is smaller than the dimensionality of the original input space. In this study an autoencoder is used[17] as a dimension reduction method and compared with principal component analysis[16] and traditional feature selection algorithms listed in section 2.5.

2.6.1. *Autoencoders*

The autoencoder (AE), a derivative of artificial neural networks, was first proposed by the Hinton and PDB groups in the 1990s.[17] In 2016, it became one of the main topics in machine learning when the deep learning architecture became more popular.[33] The autoencoder is a fully connected artificial neural network consisting of three layers: input layer, the hidden layer and the output layer. The number of neurons in the input and output layers are the same as the number of features in the dataset. The number of neurons in the hidden layer can be selected as desired, which is an important hyper-parameter that affects the performance of the network. An autoencoder does not need any class labels because it uses the input data as the output data. Therefore it is an unsupervised learning method. The network determines the optimal weight values using the backpropagation algorithm during training to match the input data to itself at the output layer with minimal loss. For this reason, the method is also referred to as the backpropagation algorithm without a teacher.[34] If there are fewer neurons in the middle layer than the output and input layers, the reduced data is derived from the middle layer. The forward propagation from one layer to the next is formulated in eq. 1.

$$y_j = f\left(\sum_{i=1}^{n} x_i \times w_{ji}\right) + b \quad (1)$$

In eq. 1, x_i represents the value of the i^{th} input in the current layer, y_j represents the value of j^{th} output in the next layer, w_{ji} represents the weight that connects the x_i to y_j, n represents the number of neurons in the current layer, b represents the bias value, and f represents the activation function (gauss, sigmoid, softmax etc.). During model training the weights are updated to minimize the difference between the actual values and the output values expressed in equation 2.

$$\min(\sum_{i=1}^{n}(y_j - y_j')^2) \quad (2)$$

where y_j represents the actual value and y_j' represents the value that is produced by the network.

By connecting several auto encoders one after another, it is possible to derive a deep autoencoder (deepAE) as shown in fig. 3. In this model, the values obtained from the hidden layer of the first autoencoder are connected to the input layer of the second autoencoder. In deep auto encoders, each autoencoder model is trained one after another. Standard autoencoder reduces the data in one step. Hence either the dimension reduces suddenly or the reduction is little. In this case, the deep auto-encoders can be used to reduce data to lower dimensions gradually, which enables more complex datasets to be separated. This is the most important advantage of deep auto encoders. In each autoencoder model, weights that connect neurons in input layer to neurons in hidden layer are called the encoder weights, and weights that connect neurons in hidden layer to neurons in output layer are called the decoder weights. After training, data with reduced number of dimensions can be obtained by forward propagating the samples through encoder weights. The aim of the decoder weights is to reproduce data in the original dimension. However new data produced at the output layer may not always be exactly the same as the original data if the encoding operation is lossy.

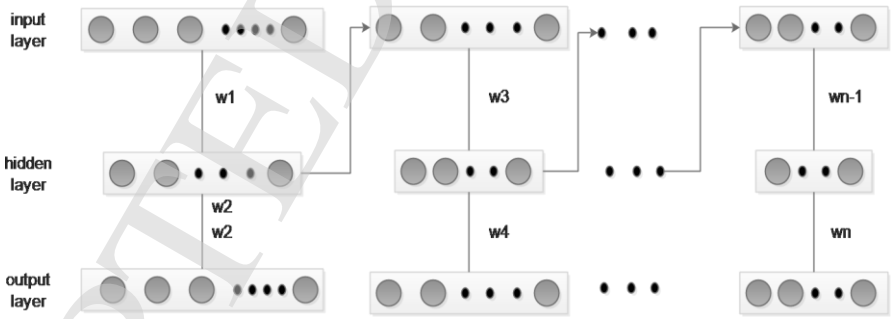

Fig. 3. Deep autoencoder architecture

In the present study the autoencoder model with one hidden layer is employed only because the accuracy of deep autoencoder model was not better than the standard auto-encoder in our preliminary experiments (results not shown).

3. Experiment and analysis

In this study, the autoencoder model explained in Section 2.6.1 is compared with traditional feature selection and dimension reduction techniques as well as with the model trained by the original dataset. For protein secondary structure prediction, the CB513 benchmark, which is originally produced by Cuff and Barton[35] and EVAset[36] benchmark are used. For solvent accessibility prediction, only the EVAset benchmark is used.[36] A one-versus-one support vector machine implemented by libSVM[37] is employed as the classifier for train-test experiments. The feature selection experiments are implemented on the Weka platform[38], the PCA is implemented in Python using the pca library[39] and the autoencoder is implemented in MATLAB.[40] Accuracy[41], SOV[42], and Matthew's correlation coefficient (MCC)[43] are used as the accuracy measures.

A 7-fold cross-validation experiment is performed on CB513 and a 10-fold cross-validation on EVAset to assess the prediction accuracy of the methods. Proteins are randomly assigned to train and test sets for each fold. Then, from each train set, 10% of the proteins are chosen randomly to form a validation set and the rest is saved as the train-set-for-optimization (i.e. model optimization or to optimize the number of dimensions). This set contains approximately 90% of the proteins in the original train set of the cross-validation. To further reduce the sample size and speed up the optimizations in EVAset, each train-set-for-optimization is further reduced by selecting 20% of the proteins randomly. Similarly, 50% of proteins are selected randomly from each validation set of EVAset benchmark. As a result, 4 different datasets are produced for each fold: train set, test set, train-set-for-optimization and validation-set-for-optimization.

True secondary structure and solvent accessibility labels of proteins in CB513 and EVAset are computed by the DSSP program[26] starting from the 3D coordinate information available in PDB. Then for each protein, PSSM and structural profiles are extracted using PSI-BLAST and HHblits as explained in Section 2.3. In the next step, the secondary structure and solvent accessibility classes are predicted using the first phase of the DSPRED method (before the SVM phase) and a total of three distributions are obtained as described in Section 2.4.1. As a result, three distributions with length L×3 for secondary structure and L×2 for solvent accessibility and two PSSM matrixes of length L×20 are obtained for each dataset, where L represents the number of amino acids in target protein. To extract features of the SVM classifier, a symmetric window of size 11 is chosen around each amino acid for CB513 and a window of size 19 for EVAset. As a result of this procedure, the number of features in each feature category, dataset and prediction problem are summarized in Table 1. In CB513, there are a total of 84119 amino acid samples and in EVAset there are 584595 amino acids.

Table 1. Number of features for each feature category, dataset and prediction problem

Problem	Dataset	PSIBLAST PSSM	HHMAKE PSSM	Dist. 1	Dist. 2	Dist. 3	Total
PSSP	CB513	220	220	33	33	33	539

PSSP	EVAset	380	380	57	57	57	931
SA	EVAset	380	380	38	38	38	874

In the third phase, for each fold of cross validation a one-vs-one SVM is trained on each train set using original feature sets (539 features for CB513, 931 features for EVAset for secondary structure prediction and 874 features for solvent accessibility prediction). Gamma parameter of the SVM is set to 0.00781 and C parameter to 1, which were optimized before by Aydin et al. for CB513.[2] Then predictions are computed on test sets.

In the fourth phase, ranker feature selection techniques (Chi-square, Information Gain, and Gain Ratio) are applied on the each train-set-for-optimization separately and features are sorted according to the calculated rank values. For each ranker method, a wrapper approach is used, in which an SVM is first trained using the feature with the best rank value only, and tested on the validation set that has the same feature. Then, other features added one by one into the datasets according to the rank order, and train and test steps are repeated for these new sets. Finally, the feature set that gives the best prediction accuracy on validation-set-for-optimization is found. Once feature selection is performed on train-set-for-optimization, the same features are selected in the corresponding validation-set-for-optimization. Feature selection steps described above are repeated for each fold of the cross-validation experiment. After the attributes are selected, for each fold of the cross-validation experiment, an SVM is trained on the original train set and class labels of test set are predicted.

In principal component analysis method the number of dimensions is increased from 5 to 535 with increments of 5, and a one-versus-one SVM model is trained and tested on each train-set-for-optimization and validation-set-for-optimization, respectively. After optimizing the number of dimensions, principal component analysis is performed on train and test datasets. Finally, for each fold, a one-versus-one SVM model is trained and tested by using the reduced versions of the original train and test sets.

In the autoencoder model, the number of hidden neurons, which gives the dimension of the reduced dataset, is increased from 75 to 525 with increments of 25. Maximum epoch number is set to 1000, L2WeightRegularization parameter to 0.004, SparsityRegularization parameter to 4, SparsityProportion parameter to 0.15 and scaleData parameter to false. As in other methods, autoencoder is applied on train-sets-for-optimization and validation-sets-for-optimization. After finding the optimum number of dimensions, a one-versus-one SVM is trained and tested on the reduced versions of original train and test sets.

Mean Q_3 accuracy (Acc), mean SOV score, mean dimension (Dim), standard deviation (SD) of difference between dimension of each fold and Matthew's correlation coefficient (MCC) for 'H', 'E', 'L' are shown in table 2 for secondary structure prediction on CB513 and in table 3 for secondary structure prediction on EVAset. The means are computed as the averages across cross-validation folds. Mean accuracy (Acc), mean SOV score, mean number of dimensions (Dim), standard deviation of difference between dimension of

each fold and mean Matthew's correlation coefficient (MCC) are shown in table 4 for solvent accessibility prediction on EVAset.

Table 2. Accuracy measures of dimension reduction methods, average number of dimensions and standard deviation values for secondary structure prediction evaluated by 7-fold cross validation experiment on CB513

Methods	**Dim**	**SD**	**MCC 'H'**	**MCC 'E'**	**MCC 'L'**	**SOV**	**ACC**
Original Dim.	539	0	0.73	0.67	0.80	79.79	82.78
X^2	365.14	76.72	0.74	0.67	0.80	80.32	81.99
IG	357.57	66.61	0.74	0.67	0.81	80.25	82.98
GR	320.86	75.85	0.74	0.67	0.80	80.35	82.96
MRMR	22.00	0.81	0.73	0.65	0.79	79.29	81.99
CFS-ge	239.71	15.71	0.73	0.66	0.80	79.66	82.53
CFS-gr	22.14	1.34	0.73	0.65	0.79	79.30	82.05
CFS-bf	22.86	1.34	0.73	0.65	0.79	79.07	82.04
PCA	504.71	11.33	0.72	0.66	0.80	79.46	82.21
Autoencoder	307.15	31.33	0.74	0.67	0.81	80.35	82.84

Table 3. Accuracy measures of dimension reduction methods, average number of dimensions and standard deviation values for secondary structure prediction evaluated by 10-fold cross validation experiment on EVAset

Methods	**Dim**	**SD**	**MCC 'H'**	**MCC 'L'**	**MCC 'E'**	**SOV**	**ACC**
Original Dim.	931	0	0.76	0.68	0.81	80.71	83.86
X^2	243.90	143.46	0.75	0.67	0.81	80.53	83.36
IG	248.60	141.71	0.75	0.67	0.81	80.55	83.36
GR	231.80	143.34	0.75	0.67	0.81	80.55	83.33
MRMR	25.20	1.32	0.74	0.66	0.80	79.53	82.57
CFS-ge	434.80	15.66	0.75	0.68	0.81	80.26	83.56
CFS-gr	29.8	2.201	0.74	0.66	0.80	79.55	82.63
CFS-bf	32	1.89	0.74	0.66	0.80	79.54	82.65
PCA	527	73.79	0.74	0.67	0.81	80.39	83.31
Autoencoder	427.5	24.86	0.75	0.67	0.81	80.32	83.05

Table 4. Accuracy measures of dimension reduction methods, average number of dimensions and standard deviation values for solvent accessibility evaluated by 10-fold cross validation experiment on EVAset

Methods	**Dim**	**SD**	**MCC**	**SOV**	**ACC**
Original Dim.	874	0	0.61	56.16	80.45
X^2	482.00	68.61	0.60	54.81	80.24
IG	510.00	62.18	0.60	54.81	80.26
GR	522.00	81.62	0.60	55.01	80.16

MRMR	20.20	1.14	0.56	51.52	77.96
CFS-ge	334.4	28.91	0.60	54.39	79.89
CFS-gr	21.20	1.40	0.56	51.51	77.97
CFS-bf	21.10	1.37	0.56	51.49	77.97
PCA	712.00	155.33	0.61	55.80	80.34
Autoencoder	402.50	14.19	0.60	54.81	80.21

Based on these results it can be observed that incorporating structural profiles into the DSPRED method improved the accuracy of secondary structure prediction by 2.5% on a difficult benchmark (CB513) and on a difficult setting, in which distant templates are used only (see Aydin et al.[2]). Furthermore as shown in experiment results, feature selection and dimension reduction algorithms can be used to reduce the number of dimensions considerably for protein secondary structure and solvent accessibility prediction. For CB513, IG obtained the best overall accuracy (Q3) value of 82.98% and CFS-MRMR method obtained the highest reduction with the mean number of dimensions equal to 22. The best overall accuracy (Q3) for secondary structure prediction on EVAset is obtained when all features are used. This is followed by CFS-Genetic search algorithm. The highest reduction in the number of dimensions is achieved by CFS-MRMR with the mean number of dimensions across the 10 folds equals to 25.20. The best accuracy for solvent accessibility prediction on EVAset is obtained when all features are used. This is followed by PCA algorithm and the highest reduction in number of dimensions is achieved by CFS-MRMR algorithm with the mean number of dimensions across the 10 folds equals to 20.20.

Table 5 summarizes the running times for PSSP and Table 6 summarizes the running times for SA of each model in a single CPU core. The train-test experiments are performed on a 2.6 GHz CPU. For the ranking and optimization steps, autoencoder simulations are performed on a 3.5 GHz CPU and all the remaining methods on a 2.6Ghz CPU. In all experiments RAM capacity is set 10 GB.

Table 5. Total running times for model optimization, training and testing for PSSP evaluated by cross validation experiments

Methods	**Dataset**	**Ranking**	**Optimization**	**Train-Test**
Original Dimension	CB513	------	------	2 days
Ranking Methods	CB513	5 hours	2 days	15 hours
CFS-Genetic	CB513	------	140 hours	12 hours
Other FS	CB513	------	2 days	5 hours
PCA	CB513	------	7 days	2 days
Autoencoder	CB513	------	6 days	14 hours
Original Dimension	EVAset	------	------	60 days
Ranking Methods	EVAset	4 days	250 days	19 days
CFS-Genetic	EVAset	------	30 days	32 days
Other FS	EVAset	------	11 days	3 days

PCA	EVAset	------	35 days	35 days
Autoencoder	EVAset	------	35 days	32 days

Table 6. Total running time for model optimization, training and testing for SA evaluated by cross validation experiments

Methods	**Dataset**	**Ranking**	**Optimization**	**Train-Test**
Original Dimension	EVAset	------	------	55 days
Ranking Methods	EVAset	4 days	195 days	32 days
CFS-Genetic	EVAset	------	26 days	3 days
Other FS	EVAset	------	10 days	3 days
PCA	EVAset	------	32 days	43 days
Autoencoder	EVAset	------	31 days	29 days

The fastest algorithms (including optimization, selection and classification) are CFS-MRMR, CFS-Best First and CFS-Greedy and the slowest algorithms are ranker feature selection algorithms for both datasets and both problems. Note that the hyper-parameters of the genetic algorithm are not optimized in this study.

To summarize, the best performing dimension reduction methods in terms of prediction accuracy are IG, GR, CFS-ge and PCA for secondary structure prediction. Among those CFS-ge provides a good compromise between accuracy and number of dimensions reducing the dimensions around 50% without sacrificing from the accuracy considerably. This has the potential of improving the model training time of the second stage of DSPRED significantly by at least four folds due to the quadratic optimization involved in SVM. The most accurate methods for solvent accessibility prediction are obtained as chi-square, IG, GR, PCA and autoencoder. Among those autoencoder has a good compromise between accuracy and the number of dimensions reducing the dimensions around 50% without sacrificing from the accuracy considerably. Regarding the compression ratio the best results are obtained by CFS-MRMR, CFS-gr and CFS-bf (dimensions are reduced to 20-30 features only) both for secondary structure and solvent accessibility prediction at the expense of a reduction in accuracy by 0.7-1.0% in secondary structure prediction and by 2.5% in solvent accessibility prediction.

Conclusions

In this study, we employ an autoencoder for dimension reduction and compare it with the traditional feature selection and dimension reduction techniques in protein secondary structure and solvent accessibility prediction on two benchmark datasets. In addition we compare the accuracy obtained after dimension reduction with the accuracy obtained using the original feature set. As the classification method we use support vector machine as the second classifier of a two-stage predictor. As a result, feature selection and dimension reduction techniques achieved similar success rates compared to the accuracy

obtained by the original feature set. They can be useful for protein structure prediction because they have the potential to decrease the number of dimensions considerably. For each feature selection and projection algorithm, the classification using the reduced feature sets is significantly faster than classification computed in the original input space. In addition, some feature selection algorithms achieve slightly higher accuracy than the models trained using the original feature set. The autoencoder model has similar success rate to the other models though it can be more ameliorative because of its several parameters. As a future work, deep learners such as deep belief networks can also be applied for dimension reduction to predict 1D structure of proteins. Furthermore, new and informative feature sets can be incorporated without reducing the computational speed of the models, which will eventually improve the accuracy of protein structure prediction.

Acknowledgments

This work is supported by grant 113E550 from 3501 TUBITAK National Young Researchers Career Award.

References

1. Hua S, Sun Z, *A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach1 1Edited by B. Holland*, J. Mol. Biol., vol. 308, no. 2, pp. 397–407, Apr. 2001.
2. Aydin Z, Singh A, Bilmes J, Noble WS, *Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure*, BMC Bioinformatics, vol. 12, p. 154, May 2011.
3. Huang YF, Chen SY, Protein secondary structure prediction based on physicochemical features and PSSM by SVM, *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2013, pp. 9–15.
4. Wang Y, Cheng J, Liu Y, Chen Y, Prediction of protein secondary structure using support vector machine with PSSM profiles, *IEEE Information Technology, Networking, Electronic and Automation Control Conference*, 2016, pp. 502–505.
5. Pugalenthi G, Kandaswamy GK, Chou KC, Vivekanandan S, Kolatkar P, *RSARF: Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method*, Protein Pept. Lett., vol. 19, no. 1, pp. 50–56, Jan. 2012.
6. Joo K, Lee SJ, Lee J, *Sann: Solvent accessibility prediction of proteins by nearest neighbor method*, Proteins Struct. Funct. Bioinforma., vol. 80, no. 7, pp. 1791–1797, Jul. 2012.
7. Adamczak R, Porollo A, Meller J, *Accurate prediction of solvent accessibility using neural networks–based regression*, Proteins Struct. Funct. Bioinforma., vol. 56, no. 4, pp. 753–767, Sep. 2004.
8. Faraggi E, Xue B, Zhou Y, *Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network*, Proteins Struct. Funct. Bioinforma., vol. 74, no. 4, pp. 847–856, Mar. 2009.
9. Mirabello C, Pollastri G, *Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility*, Bioinformatics Applications Note, pp. 2056-2058, 2013.
10. Li D, Li T, Cong P, Xiong W, Sun J, *A novel structural position-specific scoring matrix for the prediction of protein secondary structures*, Bioinformatics, pp. 32-39, 2012.

11. Pollastri G, Martin AJM, Mooney C, Vullo A, *Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information*, BMC Bioinformatics, 2007.
12. Aydin Z, Baker D, Noble WS, Constructing structural profiles for protein torsion angle prediction, *6th International Conference on Bioinformatics Models*, Methods and Algorithms, BIOINFORMATICS 2015, 2015.
13. Han J, Pei J, Kamber M, *Data Mining: Concepts and Techniques*. Elsevier, 2011.
14. Li Z, Wang J, Zhang S, Zhang Q, Wu W, *A new hybrid coding for protein secondary structure prediction based on primary structure similarity*, Gene, vol. 618, no. Supplement C, pp. 8–13, Jun. 2017.
15. Adamczak R, Dimensionality reduction of PSSM matrix and its influence on secondary structure and relative solvent accessibility predictions, *World Academy of Science*, Engineering and Technology, 2009.
16. Wold S, Esbensen K, Geladi P, *Principal component analysis*, Chemom. Intell. Lab. Syst., vol. 2, no. 1, pp. 37–52, Aug. 1987.
17. Hinton GE, Revow M, Dayan P, Recognizing Handwritten Digits Using Mixtures of Linear Models, *7th International Conference on Neural Information Processing Systems*, 1994, pp. 1015–1022.
18. Kavzoğlu T, Şahin EK, Çölkesen İ, Heyelan Duyarlılık Analizinde Ki-Kare Testine Dayalı Faktör Seçimi, *V. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu*, 2014.
19. Ozarkar P, Patwardhan M, *Efficient Spam Classification by Appropriate Feature Selection*, Glob. J. Comput. Sci. Technol. Softw. Data Eng., 2013.
20. Jeppson KO, Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay, *IEEE J. Solid-State Circuits*, vol. 29, no. 6, pp. 646–654, Jun. 1994.
21. Ding C, Peng H, *Minimum redundancy feature selection from microarray gene expression data*, J. Bioinform. Comput. Biol., vol. 03, no. 02, pp. 185–205, Apr. 2005.
22. Hall MA, *Correlation-based feature selection for machine learning*, Diss. The University of Waikato, 1999.
23. Siedlecki W, Sklansky J, *A note on genetic algorithms for large-scale feature selection*, Pattern Recognit. Lett., vol. 10, no. 5, pp. 335–347, Nov. 1989.
24. Kwak N, Choi CH, Input feature selection for classification problems, *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.
25. Xu L, Yan P, Chang T, Best first strategy for feature selection, *9th International Conference on Pattern Recognition*, 1988, pp. 706–708 vol.2.
26. Kabsch W, Sander C, *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features*, Biopolymers, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res., vol. 25, no. 17, pp. 3389–3402, Sep. 1997.
28. Remmert M, Biegert A, Hauser A, Söding J, *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment*, Nat. Methods, vol. 9, no. 2, p. 173, Feb. 2012.
29. *RCSB Protein Data Bank - RCSB PDB*, 2017, https://www.rcsb.org/pdb/home/home.do.
30. Gormez Y, *Dimensionality Reduction for Protein Secondary Structure Prediction*, M.Sc. Thesis, Abdullah Gul University, 2017.
31. Vapnik V, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 2013.
32. Cortes C, Vapnik V, *Support-vector networks*, Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

33. Baldi P, Autoencoders, Unsupervised Learning, and Deep Architectures, *JMLR: Workshop and Conference Proceedings*, 2012.
34. Rumelhart DE, James ML, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Parallel Distrib. Process., 1986.
35. Cuff JA, Barton GJ, *Evaluation and improvement of multiple sequence methods for protein secondary structure prediction*, Proteins Struct. Funct. Bioinforma., vol. 34, no. 4, pp. 508–519, Mar. 1999.
36. Koh IY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Rost ASB, *EVA: evaluation of protein structure prediction servers*, Nucleic Acids Res., vol. 31, no. 13, pp. 3311–3315, Jul. 2003.
37. *LIBSVM -- A Library for Support Vector Machines*, https://www.csie.ntu.edu.tw/~cjlin/libsvm/.
38. *Weka* , 2017, https://weka.wikispaces.com/ .
39. *Principal Components analysis*, 2017, http://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_3d.html.
40. *trainAutoencoder*, 2017, https://www.mathworks.com/help/nnet/ref/trainautoencoder.html.
41. *Precision and recall*, 2017, https://en.wikipedia.org/wiki/Precision_and_recall.
42. Zemla A, Venclovas C, Fidelis K, Rost B, *A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment*, Proteins, vol 34, pp. 220–223, 1999.
43. Matthews BW, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*, Biochim Biophys Acta, vol. 405, no. 2, pp. 442-451, 1975.