International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:3, No:10, 2009

Dimensionality reduction of PSSM matrix and its influence on secondary structure and relative solvent accessibility predictions

Rafał Adamczak

***Abstract*—**State-of-the-art methods for secondary structure (Porter, Psi-PRED, SAM-T99sec, Sable) and solvent accessibility (Sable, ACCpro) predictions use evolutionary profiles represented by the position specific scoring matrix (PSSM). It has been demonstrated that evolutionary profiles are the most important features in the feature space for these predictions. Unfortunately applying PSSM matrix leads to high dimensional feature spaces that may create problems with parameter optimization and generalization. Several recently published suggested that applying feature extraction for the PSSM matrix may result in improvements in secondary structure predictions. However, none of the top performing methods considered here utilizes dimensionality reduction to improve generalization. In the present study, we used simple and fast methods for features selection (t-statistics, information gain) that allow us to decrease the dimensionality of PSSM matrix by 75% and improve generalization in the case of secondary structure prediction compared to the Sable server.

***Keywords*—**secondary structure prediction, feature selection, position specific scoring matrix

I. Introduction

PROTEIN structure prediction from the amino acid sequence is an fundamental and challenging problem in molecular biology. Stimulated by the difficulty of the overall structure prediction, computational methods for the prediction of intermediate attributes, such as secondary structure (SS)[1] [2] [3], solvent accessibility (SA) [4] [5], contact number [6], are being developed in order to facilitate protein folding simulations, structure prediction and functional annotation of important domains, motifs and individual amino acid residues. These problems are not only easier to solve (compared to the 3d structure prediction), but are also better suitable for machine learning approaches.

In order to apply machine learning algorithms, proteins (and individual amino acid residues) are typically represented as vectors. For example, Artificial Neural Networks and Support Vector Machines, which are widely used in structural bioinfomatics, require vector representation in some feature space. In fact, the choice of the representation is crucial for a successful development of accurate and robust predictors with good generalization. Many descriptors of residues (features), such as hydrophobicity, polarity, amino acid propensities etc., have been applied to distinguish, e.g., if residue is in a helix or strand in case of secondary structure prediction [7], [8] , or if it is in contact with the other residue in the case of the contact map prediction [9]. However, it has been shown that evolutionary-based representations that utilize multiple alignment and information about protein families, as encoded by the position specific scoring matrix (PSSM) [10], yield the highest accuracies in secondary structure and solvent accessibility problems [1], [6]. In the PSSM, each amino acid residue ("position" in the sequence) is represented by 20 numbers (20 possible amino acid substitutions) that reflect frequencies of substitutions observed at this position in a protein family. PSSM scores are positive, indicate that given amino acids substitution occurs more frequently in the alignment that expect by chance, and negative, substitution occurs less frequently then expected. Multiple alignments and the resulting PSSMs are typically computed iteratively, e.g., using the Psi-BLAST program [11].

From the machine learning point of view there is one disadvantage of using evolutionary profiles in data vectors – size of the PSSM matrix, in other words number of features. The more features in the training vector the bigger size of the training set, to fully cover feature space, must be. Models that need to be trained are minimizing some function that depends on the number of features and high dimensional functions have potential to be much more complicated then the lower one so its harder to find good local minimum. This may lead to poor generalization and of course long training time. Dimensionality reduction is the solution for this problem and also may help to understand the data, it is possible to find features that are most informative or most important for particular problem.

So far most prediction methods, were PSSM matrix has been used, are using raw PSSM matrix without dimensionality reduction. There are only few examples of dimensionality reduction of PSSM matrix [12], [13] where PCA (principle component analysis) [14] and C-NLPCA (cascaded non linear components analysis) has been used. Both papers reports improvements in secondary structure prediction on database developed by Cuff and Barton CB396 [15]

This article presents application of features extraction, features selection and combination of these two methods on PSSM matrix and its influence on secondary structure and solvent accessibility predictions. For features extraction PCA has been used, features selection is based on information gain [16] and t-statistics.

II. Materials and Methods

A. Training and control sets

Training set has been build using the same protein structures as were used to develop Sable [4] [17] method. In that case representative and non-redundant set of protein chains was created based on Pfam (Protein Families) database, version 6.6. After careful preprocessing of Pfam families and its representative PDB structures 860 PDB structures (about 210000 residues) with no homology between each other were selected.

For control sets we used the same sets as it was used for Sable server evaluation. There are 603 protein chains (with 143000 residues) with no homology to proteins included in the training that are grouped into 4 datasets referred to as S156 (156 structures submitted to PDB from January through March of 2002), S135 (135 structures submitted from April through June), S163 (163 structures submitted from July through September), and S149 (149 structures submitted from October through December of 2002). The list of protein structures in the training and all control sets can be downloaded from http://sable.cchmc.org.

B. Feature space

For secondary structure and solvent accessibility we used the same feature space as it was used in data generated to train

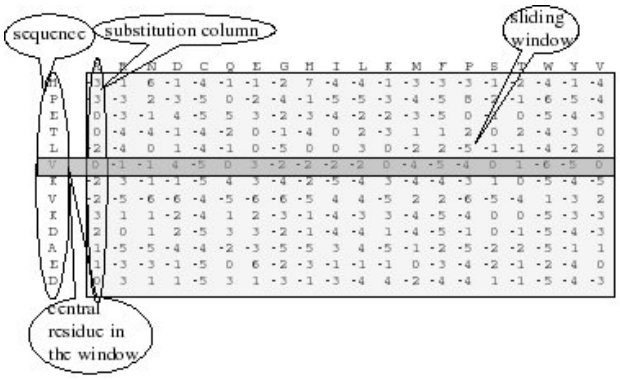

Fig. 1: Sliding window on the PSSM matrix.

Sable server. The local structural environment and evolutionary context of each residue is characterized by a

sliding window of 11 amino acids, with the residue of interest at position 6. The most important part of the feature space is evolutionary information represented in the form of position-specific scoring matrix (PSSM) generated by PSI-BLAST program (version 2.2.9 with default options) on nr database (02.12.2007 with 5678482 sequences). For the window of the size 11 there is 220 features obtained from PSSM matrix, since each amino acid in the window is represented in the PSSM matrix by 20 columns of substitution scores. We tried also other window sizes but 11 gave best results for secondary structure (SS) and solvent accessibility (SA) predictions

Each amino acid in the window is characterized by entropy, weighted hydrophobicity and volume. Probabilities, needed to calculate entropy, and weights are frequencies of amino acids obtained from multiple alignment, so these 33 features also include evolutionary information.

$$V = \sum_{i=1}^{20} p_i V_i$$

where p_i probability of occurring amino acid *i*, V_i volume (hydrophobicity) of *i* amino acid.

Central residue in the window and its two immediate neighbors are represented by binary vector of length 5, value 1 in this vector indicates the presence of amino acids belonging to 1 of the 5 groups with distinct SS propensities: {A, E, L}, {V,I} , {S, N}, {P}, {G}. There is one more feature indicating the presence of cysteine residues in the window. Altogether input space for solvent accessibility and secondary structure prediction consists of 269 features.

For secondary structure prediction we used two steps of predictions that have been introduced in PHD [18] called: sequence to structure and structure to structure. Characterized feature space has been applied for the first step of prediction, the second one is created using only predictions of the first step. As in the first step training vectors are created by sliding window of size 11 on predicted secondary structures. Also architecture of networks (number of hidden nodes) used in structure to structure phase was the same as for sequence to structure networks.

C. Training and testing protocol

For secondary structure and solvent accessibility prediction neural networks, generated and trained by Stuttgart Neural Network Simulator (SNNS) [19], have been applied. In all cases Rprop [20] algorithm with default parameters has been used. The best architecture, number of hidden nodes in hidden layer (we used networks with only one hidden layer), has been determined in 10 fold cross validation. For both problems networks with 30 hidden nodes have been selected. Architecture optimization has been made only for training set with full feature space, for training set where feature space has been reduced we used the same number of hidden nodes as in the optimized architecture.

For each fold of 10 fold cross validation 3 sets have been produced: training set 80% of full set, validation set and internal testing set 10%. Training in every fold was performed as long as accuracy on validation set did not drop, what was checked every 10 epochs. Cross validation accuracy is obtained by averaging results obtained on test set by selected networks.

D. Relative Solvent Accessibility and Secondary Structure

Relative solvent accessibility (RSA) is defined as the ratio of the solvent-exposed surface area of that residue observed in a given structure, denoted as SA_i, and the maximum obtainable value of the solvent-exposed surface area for this amino acid, denoted as MSA_i:

$$RSA_i = 100 * \frac{SA_i}{MSA_i}$$

Thus, RSA_i adopts values between 0% and 100%, with% corresponding to a fully buried and 100% to a fully accessible residue, respectively. For convenience values of RSA has been rescaled to [0,1].

Following training protocol of Sable server we changed standard cost function used to train neural network (SSE sum squared error) to weighted SSE.

$$wSSE = \sum_i \alpha(o_i)(y_i(z) - o_i)^2$$

where $y_i(z)$ is the predicted value for the ith input vector given the parameters of the network (weights and biases) z, and o_i represents the observed real-value RSAs that are imposed in the training, weights $\alpha(o_i)$ are defined in [4] and they account for naturally occurring variability in terms of RSA in families of homologous structures.

Secondary structure (SS) and solvent accessible (*SA*) for selected proteins have been generated using Dictionary of Protein Secondary Structure (DSSP) program [21]. Since the DSSP program assigns each residue to 1 of 8 distinct secondary structure classes, the following conversion from 8 to 3 classes was applied: {G, I, H} to H, {B, E} to E, and {T, S, and "other"} to C, where H denotes helix, E denotes - strand, and C denotes coil.

E.Dimensionality reduction

There are two main approaches for dimensionality reduction: feature selection and feature extraction. Feature selection algorithms find the most relevant subset of original features that help to build robust learning model, and better understand the data. Feature selection methods can be divided into two categories: ranking features and subset selection. The first group of methods provides information on how important certain subsets of features are for the particular task at hand. The subset selection however is using some search algorithms to find the best subset of features. Each subset is evaluated using some machine learning model. Usually these class of methods give better results but disadvantage is high computational cost, which is an important consideration in our case. We applied two fast ranking features methods: t-statistics and information gain. Number of *m* features that leads to best generalization has been determined in cross validation.

The goal of feature extraction is to find mapping from high dimensional space to lower one preserving at the same time most of the information and structure of the original space. It always leads to completely different feature space and as a consequence understanding of the original data based on original features is considerable more difficult. For the feature extraction PCA has been used.

1)Information gain

For data $X=\{x_i;i=1...m\}$ divided into k classes $C=\{c_j;j=1..k\}$, where each vector x_i in the data is represented by n features $F_l=\{f_{i,}i=1...s_l\}$, where $l=1...n$, s_l is the number of different values in the feature F_l, information gain for feature F_l is defined by:

$$I(F_l) = H(C) - H(C \mid F_l)$$

where $H(C)$ is the entropy of classes C in the data:

$$H(C) = \sum_{i=1}^{k} p(c_i)\log(p(c_i))$$

where $p(c_i)$, is the probability of class c_i occurring

$$H(C \mid F_l) = \sum_{i=1}^{n} p(f_i^l)\sum_{j=1}^{k} p(c_j \mid f_i^l)\log(p(c_j \mid f_i^l)$$

where $p(f_i^l)$ is the probability occurring value f_i^l in the feature F_l and $p(c_j \mid f_i^l)$ is the conditional probability that class c_j occurs giving that f_i^l has occurred.

2)T-statistics

T-statistics for feature F and data divided into two classes (plus, minus) is defined as follows:

$$Tstat(F) = \frac{\mid \mu_{plus} - \mu_{\min us} \mid}{\sqrt{\frac{\sigma_{plus}^2}{n_{plus}} + \frac{\sigma_{\min us}^2}{n_{\min us}}}}$$

where μ_{plus} mean value calculated for vector from class plus, μ_{minus} mean value for vectors from class minus, σ_{plus} standard deviation for data in *plus* class, n_{plus} number of vectors in *plus* class, and σ_{minus} standard deviation for data in *minus* class, n_{minus} number of vectors in *minus* class. In our case we have multi class problems, *Tstat* has been calculated for each of the class separately (particular class versus others) and summed over all classes.

3)Principle component analysis

PCA is a linear transform widely used in data analysis and compression. It transforms input data *X (*matrix *m*x*n,* where *n* number of features, *m* number of vectors*)* into new feature space that is build based on eigenvectors of covariance matrix *C* derived from the data *X*.

$$Y = XP$$

where, *P* *n*x*n* eigenvectors of covariance matrix *C* (each eigenvector is in the column of *P*)*, Y* representation of data *X* in new coordinate system.

This projection assures that the top components (associated with the largest eigenvalues) capture most of the variance of the input data. Level of variance for each dimension of the transformed system is determined by eigenvalue of covariance matrix *C*. Removing dimensions with the smallest eigenvalues (smallest contribution to variance) leads to dimensionality reductions. Level of dimensionality reduction depends on the problem and must be found empirically.

III. Results

PSSM applied for solvent accessibility and secondary structure prediction is represented by sliding window of sequence amino acids what leads in our case to 220 features. To check similarity between features we calculated Pearson correlation coefficient for each feature pair, derived from PSSM, in the training data. Although there is 220 features for correlation analysis we used only 20 features from the middle of the window, because correlations between features from different positions in the window (each position in the window is represented by 20 features) are very small (on the level of 0.1). The highest correlation that we observed was 0.82 between feature that is representing substitution by valine (V) and isoleucine (I), the next one is 0.74 between isoleucine and leucine (L). Correlations between features are represented in the right panel in , low correlation (less the 0.5) is marked by black color, all others are marked by gray scale. The higher lightness of the gray scale the higher correlation, correlation 1 is marked by white color and it appears only on the diagonal position. There are eight features that are not highly correlated to any other feature A, C, G, H, P, S, T, W. To see overall similarity average correlation over all features in PSSM matrix has been calculated and its profile is presented on the left panel in . The smallest similarity is observed for alanine, small values have also proline, threonine, tryptophan and cysteine. Because features in PSSM are so highly correlated dimensionality reduction based on PCA, that leads to uncorrelated feature space, should work well on this data. We test this hypothesis specifically in the case of secondary structure prediction [12].

To assess the importance of each of the 220 features for secondary structure and solvent accessibility predictions we calculated information gain, and the results are presented in Figure 3. Since information gain can be calculated for classification problems only, we converted relative solvent accessibility values into 10 classes, dividing the overall range of RSAa from [0,1] into 10 equally sized bins with the width of 0.1. We tested other number of classes using bins with 0.05 and 0.2 width, concluding that the results were largely insensitive to this choice.

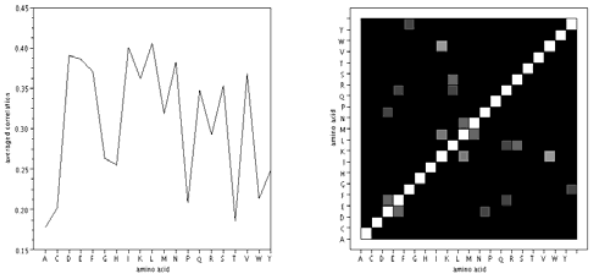

Fig. 2 Left panel represents features profile of average correlation coefficients for particular amino acids over all columns in PSSM matrix, the right panel presents correlation matrix for PSSM columns, black color means regions with low correlations (less then 0.5), lightness of the gray scale higher correlations, the highest correlations is depicted by white color.

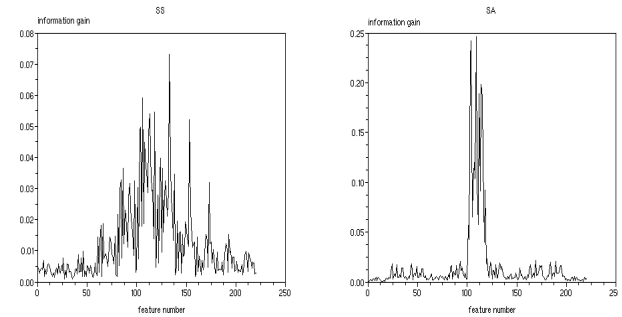

Fig. 3: Plots represents value of information gain for PSSM features obtained for window 11, left panel for secondary structure classes, right for relative solvent accessibility mapped into 10 classes.

Both graphs in Figure 3 confirms well known facts that solvent accessibility prediction depends mostly on the signal from the middle residue in the window (middle 20 features) [22], and in case of secondary structure signal is spread in the neighborhood of the middle amino acid. There is one characteristic that can be seen on both graphs - periodicity for every 20 features. Because of that we created two different strategies for feature selection: based on scoring for all 220 features and based on sum of all scores for each amino acid substitution over positions in the window of particular amino acid what irrespective of the window size gives always 20 features (these features we call substitution columns, graphical representation of substitution columns is depicted on the). For scoring substitution columns we used information gain and t-statistics. Profiles of PSSM columns for secondary structure and relative solvent accessibility are presented on Figure 4 and Figure 5 accordingly.

A. Secondary structure prediction

Feature selection of PSSM matrix for sequence to structure step has been made based on profiles depicted on Figure 3 and

Figure 4 in case of substitution columns. Figure 4 presents averaged over window t-statistics profile (left panel) and information gain profile (right panel). Information gain and t-statistics profiles are very similar, in both cases the highest values is getting by proline and glycine substitution column, the worst six columns, with slightly different order, is the same (W, Y, C, F, H, T). Because of that we decided to use only information gain for further calculations.

We checked importance of the best substitution columns by removing one column from the PSSM matrix. The highest accuracy dropping in CV was observed for proline - 2%, alanine – 1% and valine – 0.5%. What is interesting we did not notice accuracy dropping in case of glycine, although it is not highly correlated to the other columns. Removing the worst features, we were able to decrease by 11 (W, Y, C, F, H, T, S, M, R, I, N) number of columns in PSSM matrix having better accuracy in CV compared to predictor without feature selection. Next feature according to information gain was L but removing it decreased accuracy in CV. We checked all remaining features and finally reduced PSSM matrix to 5 substitution columns P, A, V, E, L (predictor S1_PAVEL). This means 55 features out of 220 and reduction by 75%. To bind founded substitution columns with secondary structure we looked at positive values in PSSM matrices generated for each query in training set. We observed that substitution by P appears mostly in coils – 65% of all occurrence of positive values for this class, by A in helices - 53%, by V in beta strands 43%, by E in helices - 51% and by L in helices – 48%. We have to add that V is the only one amino acid substitution column that has higher occurrence of positive values in beta

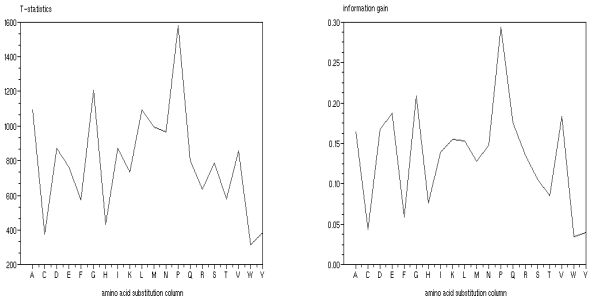

Fig. 4 Left and right panel presents amino acid profiles calculated by using t-statistics and information gain values for each amino acid substitution column and secondary structure classes.

strands compared to other classes.

The second strategy of feature selection was to use regular features. Because we were able to decrease dimensionality of PSSM matrix to 5 substitution columns, we tried to use 55 features with the best information gain. In cross validation results were much worse then using 5 substitution columns (by 0.5%).

Comparison of our predictors have been made based on Q3, percentage of residues correctly predicted in a sequence and averaged over all sequences (EVA - EValuation of Automatic protein structure prediction, like methodology [23]), and SOV measures [24]. Results on four test sets, obtained by averaging accuracy from 10 neural networks generated in cross validation, are presented in Table 1. The worst generalization is obtained for networks without feature selection (S1_SSR), the best for predictor that was using feature extraction of PSSM matrix obtained by PCA.

TABLE I SECONDARY STRUCTURE PREDICTION RESULTS FOR SEQUENCE TO STRUCTURE PHASE, OBTAINED BY AVERAGING ACCURACY FROM 10 NEURAL NETWORKS GENERATED IN CROSS VALIDATION. FOR EACH Q3 AND SOV MEASURES STANDARD DEVIATION IS GIVEN. FOLLOWING PREDICTORS HAVE BEEN COMPARED: S1_SSR – PREDICTOR WITH FULL FEATURE SPACE, S1_PAVEL - PREDICTOR WITH 5 BEST SUBSTITUTION COLUMNS, S1_SSF – PREDICTOR WITH 55 BEST FEATURES ACCORDING TO INFORMATION GAIN. PCA ABBREVIATION MEANS THAT FOR PARTICULAR FEATURE SPACE PRINCIPLE COMPONENT ANALYSIS HAS BEEN APPLIED.

	S135		S149		S156		S163	
	Q3	SOV	Q3	SOV	Q3	SOV	Q3	SOV
S1_SSR	75.9+/-0.3	72.3+/-0.4	74+/-0.2	70+/-0.3	74.1+/-0.3	70.5+/-0.4	75.1+/-0.2	70.7+/-0.2
S1_SSR_PCA	77.1+/-0.2	73.2+/-0.5	75+/-0.2	70.6+/-0.3	75.4+/-0.4	71.6+/-0.4	76.1+/-0.2	72+/-0.3
S1_PAVEL	76.5+/-0.3	72.6+/-0.4	74.5+/-0.2	70+/-0.2	75.0+/-0.3	71.3+/-0.2	75.5+/-0.2	71.3+/-0.4
S1_PAVEL_PCA	76.8+/-0.2	73.1+/-0.3	74.7+/-0.2	70+/-0.3	75.3+/-0.3	71.3+/-0.6	75.7+/-0.2	71.5+/-0.5
S1_SSF	76.6+/-0.1	72+/-0.3	74.4+/-0.1	70+/-0.4	74.7+/-0.3	70+/-0.4	75.5+/-0.2	71.1+/-0.3

Similar results have been obtained for structure to structure networks, predictors with applied dimensionality reduction are better and PCA is giving the best results. Situation changed for our final predictors, committee of networks, although predictors with dimensionality reduction are still better differences between dimensionality reduction methods vanished. Committee of structure to structure networks, derived from sequence to structure networks, are obtained by summing probabilities for each of the class from all committee networks. Class with the highest probability was the output of the committee. Because results for all dimensionality reduction methods are the same we decided to use, for further comparison, the one with the smallest number of substitution columns in PSSM matrix - S2_PAVEL that was build on the top of S1_PAVEL method mentioned above. S2_PAVEL consists of 10 networks obtained from cross validation. We compared results of S2_PAVEL predictor with one of the state of the art method – Sable that has been evaluated by EVA web based server. Sable has two secondary structure prediction methods Sable1 and Sable2, the only differences between these two methods is that in Sable2 for training structure to structure networks predicted solvent accessibility has been used, what is giving usually better results. To have fair comparison we used stand-alone version of Sable (downloaded from ftp://ftp.chmcc.org/pdi/jmeller/sable/) thanks to this we were able to use the same nr database for PSSM matrix generation for all predictors. Results in Table 2 show that S2_PAVEL predictor is much worse then Sable1 and Sable2. The strength of Sable methods comes from big diversity of networks in the committee that has been obtained

by applying different networks architectures and training algorithms. To improve generalization of our predictor we increased diversity of networks by using different datasets, applying various substitution columns in PSSM matrix, for training networks with the same architecture. We built new predictor (S2_COM) that consists of networks trained with following PSSM columns: PAVEL, PAVELIT, PAVGDI, PAGKLMN, GMTVQR, PIDVKL, AGSVLR, PHWMI, PAMSQR. It must be noted here that we did not optimized number of substitution columns that should be used to obtained best result for secondary structure prediction. We were trying to created training sets diverse as much as possible taking into account information gain and correlation between substitution columns. In most cases P, A, V has been used as the most important, we also arbitrary limited number of substitution columns between 5 and 7.

As can be seen in Table 2 increasing diversity of networks in the committee significantly improved results for S2_COM predictors, which are now comparable to Sable1. Moreover because the strategy of committee creation is very different then in Sable1, making combination of these two predictors, by taking prediction with higher probability, significantly improved results in Q3 and SOV. What is interesting Sable1+S2_COM is giving almost identical results as Sable2 particularly in Q3 measure. We tried also to create predictor trained on S1_PAVEL predictions and predicted solvent accessibility, in the same way as it is done in Sable2, but there was no improvement whatsoever. Combining S2_COM and Sable2 gave on two datasets some small improvement in Q3 and SOV measure.

TABLE II RESULTS FOR SECONDARY STRUCTURE PREDICTIONS, S2_PAVEL – PREDICTOR CONSISTS OF 10 NETWORKS ALL OF THEM HAS BEEN TRAINED ON DATA IN WHICH PSSM MATRIX WAS REPRESENTED BY 5 SUBSTITUTION COLUMNS: P, A, V, E, L, S2_COM – PREDICTOR CONSISTS OF 9 NETWORKS TRAINED WITH THE PSSM MATRIX REPRESENTED BY FOLLOWING SUBSTITUTION COLUMNS: PAVEL, PAVELIT, PAVGDI, PAGKLMN, GMTVQR, PIDVKL, AGSVLR, PHWMI, PAMSQR, SABLE1 AND SABLE2 PREDICTORS FROM SABLE SERVER, SABLE1+S2_COM AND SABLE2+S2_COM PREDICTORS OBTAINED FROM COMBINATION OF SABLE AND S2_COM PREDICTORS.

	S135		S149		S156		S163	
	Q3	SOV	Q3	SOV	Q3	SOV	Q3	SOV
S2_PAVEL	77.9	75	76.7	73.6	75.9	72.8	77.5	74.6
S2_COM	78.7	75.6	77.6	74.6	76.5	73.8	78.4	75.4
Sable1	78.6	76.2	77	74.2	77.3	75.1	78.2	75.2
SABLE1+S2_COM	79.1	76.5	77.6	74.6	77.4	75.3	78.8	75.8
Sable2	79.1	76.6	77.6	75.1	77.4	74.4	79.5	76.3
SABLE2+S2_COM	79.1	75.9	78	75	77.7	75.1	79.5	76.7

B.Solvent accessibility prediction

Profiles obtained using measure of correlation and information gain Figure 5 are very similar, even more then in secondary structure case, especially in the region of best and worst features. As in secondary structure case we decided to use only information gain. Right graph in Figure 3 shows strong signal from the middle position in the window. Removing any feature, substitution column, from this position led to dropping accuracy. We decided to keep middle position untouched and remove features from other 10 window positions. Features were removed according to the lowest value of information gain obtained as a sum over 10 positions for each amino acids. In each step we checked results in 10 fold cross validation. In this way we were able to remove 10 features W, Y, T, C, F, M, A, V, I, L. After removing features we applied also PCA. Optimal number of eigenvectors has been determined in cross validation, highest accuracy in case of predictor with full feature space (SAR) has been obtained for 50 eigenvectors and in case of predictor with selected substitution columns in PSSM (SAF) 40. Final predictors are calculated by averaging predictions from 10 networks obtained from cross validation. Evaluations were made using following measures: correlation coefficient CC, mean absolute error MAE. Results for solvent accessibility predictions for different methods are presented in Table 3.

We compared our predictors with Sable method. The differences between different predictors are very small, but there is one dataset (149) where Sable results are much better then the others. Because SAR predictor is also worst for this dataset we belief that the difference comes from more diverse neural network used by Sable.

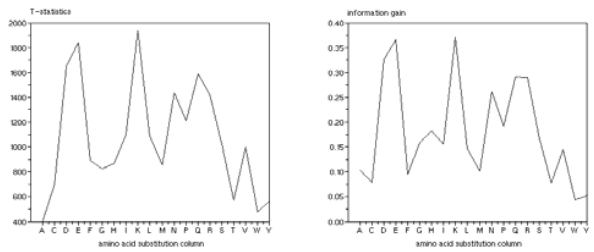

Fig. 5 Left and right panel presents amino acid profiles calculated by using t-statistics and information gain values for each amino acids and 10 solvent accessibility classes.

TABLE III COMPARISON OF SOLVENT ACCESSIBILITY PREDICTION, SAR - PREDICTOR WITH FULL FEATURE SPACE, SAF - PREDICTOR WITH 10 SUBSTITUTION COLUMNS, PCA ABBREVIATION MEANS THAT FOR PARTICULAR FEATURE SPACE PRINCIPLE COMPONENT ANALYSIS HAS BEEN APPLIED

	S135		S149		S156		S163	
	CC	MAE	CC	MAE	CC	MAE	CC	MAE
SAR	0.67	15.0	0.64	15.8	0.66	15.4	0.66	15.3
SAR_{PCA}	0.68	14.9	0.63	15.9	0.66	15.3	0.66	15.3
SAF	0.68	15.0	0.62	16.0	0.67	15.4	0.66	15.4
SAF_{PCA}	0.68	14.9	0.63	15.9	0.66	15.4	0.67	15.3
Sable	0.67	15.0	0.66	15.5	0.66	15.4	0.67	15.3

IV. CONCLUSIONS

In this paper we proposed application of feature selection to secondary structure and solvent accessibility predictions. Although proposed feature selection does not lead directly to best predictions it helped to understand the data and allowed

to change typical strategy used in building best predictors. It has been especially useful in secondary structure prediction were best methods are based on committee of predictors. The usually strategy is to create ensemble of predictors obtained by training on different datasets produced by some kind of cross validation or by using different training algorithms. The goal of this strategy is to create predictors that are highly accurate and at the same time are diverse because only in that case committee will work well. Proposed feature selection allowed to build different training sets using different substitution columns in PSSM matrix. We found that the most important substitution columns are P, A and V, moreover using additional E and L we were able to build very accurate individual predictor. Using some of those substitution columns and adding some other that are not highly correlated to the existing ones we were able to build committee, from 9 networks trained with the same training algorithm and using the same architecture, results of which are comparable to the one of the state of the art method – Sable1. Furthermore, thanks to different strategy in creating predictors for committee, combination of newly created predictor and Sable1 significantly improved results.

Feature selection for solvent accessibility problem is much harder to make, mostly because almost all information comes from middle position in the window, other positions have very small influence on solvent accessibility prediction. Nevertheless we were able to decrease dimensionality of PSSM matrix by 45% in case of feature selection and by 80% using PCA.

Acknowledgments

We thank to Jarek Meller for stimulating discussions and the BMI cluster at Cincinnati Children's for computer time on its Linux cluster.

References

[1] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices., J Mol Biol 292 : 195-202.
[2] Pollastri, G. & McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction., Bioinformatics 21 : 1719-1720.
[3] Rost, B. (2001). Review: protein secondary structure prediction continues to rise., J Struct Biol 134 : 204-218.
[4] Adamczak, R.; Porollo, A. & Meller, J. (2004). Accurate prediction of solvent accessibility using neural networks-based regression., Proteins 56 : 753-767.
[5] Pollastri, G.; Martin, A. J. M.; Mooney, C. & Vullo, A. (2007). Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information., BMC Bioinformatics 8 : 201.
[6] Pollastri, G.; Baldi, P.; Fariselli, P. & Casadio, R. (2001). Improved prediction of the number of residue contacts in proteins by recurrent neural networks., Bioinformatics 17 Suppl 1 : S234-S242.
[7] King, R. D. & Sternberg, M. J. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction., Protein Sci 5 : 2298-2310.
[8] Woodcock, S.; Mornon, J. P. & Henrissat, B. (1992). Detection of secondary structure elements in proteins by hydrophobic cluster analysis., Protein Eng 5 : 629-635.
[9] Bastolla, U.; Porto, M.; Roman, H. E. & Vendruscolo, M. (2005). Principal eigenvector of contact matrices and hydrophobicity profiles in proteins., Proteins 58 : 22-30.
[10] Gribskov, M.; McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins., Proc Natl Acad Sci U S A 84 : 4355-4358.
[11] Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs., Nucleic Acids Res 25 : 3389-3402.
[12] Melo, J. C. B.; Cavalcanti, G. D. C. & Guimaraes, K. S. (2003). PCA feature extraction for protein structure prediction, 4 : 2952-2957.
[13] Simas, G. M.; Botelho, S. S. C.; Grando, N. & Colares, R. G. (2008). Dimensional Reduction in the Protein Secondary Structure Prediction — Nonlinear Method Improvements. In: (Ed.), Innovations in Hybrid Intelligent Systems, Springer Berlin / Heidelberg.
[14] Jollife, I. T., 1986. Principle component analysis. Springer Varlag, .
[15] Cuff, J. A. & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction., Proteins 34 : 508-519.
[16] E. Hunt, J. Martin, P. S. (1966). Experiments in Induction, Academic Press, New York .
[17] Adamczak, R.; Porollo, A. & Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins., Proteins 59 : 467-475.
[18] Rost, B.; Sander, C. & Schneider, R. (1994). PHD--an automatic mail server for protein secondary structure prediction., Comput Appl Biosci 10 : 53-60.
[19] Zell, A.; Mache, N.; Hubner, R.; Mamier, G.; Vogt, M.; uwe Herrmann, K.; Schmalzl, M.; Sommer, T.; Hatzigeorgiou, A.; Doring, S.; Posselt, D.; Reczko, M. & Riedmiller, M. (1993). SNNS - Stuttgart Neural Network Simulator, .
[20] Riedmiller, M. & Braun, H. (1992). RPROP- A fast adaptive learning algorithm, .
[21] Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features., Biopolymers 22 : 2577-2637.
[22] Zemla, A.; Venclovas, C.; Fidelis, K. & Rost, B. (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment., Proteins 34 : 220-223.
[23] Eyrich, V. A.; Martí-Renom, M. A.; Przybylski, D.; Madhusudhan, M. S.; Fiser, A.; Pazos, F.; Valencia, A.; Sali, A. & Rost, B. (2001). EVA: continuous automatic evaluation of protein structure prediction servers., Bioinformatics 17 : 1242-1243.
[24] Wagner, M.; Adamczak, R.; Porollo, A. & Meller, J. (2005). Linear regression models for solvent accessibility prediction in proteins., J Comput Biol 12 : 355-369.