Biocybernetics and Biomedical Engineering
2013, Volume 33, Number 1, pp. 47–61

Feature Selection of Protein Structural Classification Using SVM Classifier

ZBIGNIEW KRAJEWSKI*, EWARYST TKACZ

Silesian University of Technology, Gliwice, Poland

Recursive feature elimination method (RFE), cross validation coefficient (CV) and accuracy of classification of test data are applied as a criterion of feature selection in order to find relevant features and to analyze their influence on classifier accuracy. Feature selection method was compared to principal component analysis (PCA) to understand the effectiveness of feature reduction. Support vector machine classifier with radial basis function (RBF) kernel is applied to find the best set of features using grid model selection and to select and assess relevant features. The best selected feature set is then analyzed and interpreted as the source of knowledge about the protein structure and biochemical properties of amino acids included in the protein domain sequence.

K e y w o r d s: pseudo amino acid composition, support vector machine, principal component analysis, recursive feature elimination, feature selection, SCOP database

1. Introduction

Reduction of data dimension to optimal feature subset makes obtaining the better accuracy of classifier possible or at least improves computational abilities. High dimensional feature set could be the reason of over-learning because of high VC dimension, leading to increase of guaranteed risk [1, 2]. Feature extraction by the aid of widely used principal component analysis (PCA) by projection into the principal components, where new features become linear combination of original features [3], obviously causes the lost of information related to interesting influence of original features on classifier accuracy [1]. PCA enables data decorrelation in order to dimension reduction in a new feature space. Covariance matrix Σ is applied determined from the observation data set as follows:

* Correspondence to: Zbigniew Krajewski, Silesian University of Technology, ul. Akademicka 16, 44-101 Gliwice, Poland, e-mail: Zbigniew.Krajewski@polsl.pl
Received 05 March 2012; accepted 14 September 2012

$$\Sigma = E[\mathbf{x}\mathbf{x}^{\mathbf{T}}]. \tag{1}$$

The aim is to separate uncorrelated components with the highest variance values. Following transformation is applied:

$$\mathbf{y} = \mathbf{T}^{\mathbf{T}}\mathbf{x}, \tag{2}$$

so as

$$\Sigma_y = \mathbf{T}^{\mathbf{T}} E[\mathbf{x}\mathbf{x}^{\mathbf{T}}]\mathbf{T} = \mathbf{T}^{\mathbf{T}}\Sigma\mathbf{T} = \mathbf{D} \tag{3}$$

where **D** is diagonal matrix and $\Sigma\mathbf{T} = \mathbf{T}\mathbf{D}$ is an eigenvector matrix condition [4].

Selection of relevant features enables to avoid PCA restriction [5]. The aim is to find original features which affect classifier accuracy together with other features and to eliminate features which remain relevant but useless [1]. Generally, feature selection methods could be assigned to one of three categories: filter, wrappers and embedded. In the filter methods, the feature rank based on information which doesn't depend on classifier is applied [6, 7]. The methods of wrappers use trained classifiers treated as "black box" as criterion of the feature set choice. The embedded methods belong to group of algorithms, which realize feature selection during classifier learning. These methods are very precise and fast but they could perform feature selection only for applied classifiers [8].

Recursive feature elimination method (RFE) is meant for nonlinear multivariate feature selection with the use of mechanism of support vector machine (SVM) classifier [9]. Feature selection is performed on the basis of influence of *w* vector on an objective function *Q* determined as Taylor series [10]. After respective approximations: diagonal, extreme and quadratic, the objective function becomes as follows:

$$\partial Q = \frac{1}{2}\sum_i \frac{\partial^2 Q}{\partial w_i^2}\partial w_i^2. \tag{4}$$

And finally for SVM classifier, the objective function is determined as [11]:

$$\begin{aligned} c_f &= \left| \|w\|^2 - \left\|w^{(-f)}\right\|^2 \right| = \\ &= \frac{1}{2}\left| \sum_{i,j=1}^{M} \alpha_i^* \alpha_j^* y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i,j=1}^{M} \alpha_i^{*(-f)} \alpha_j^{*(-f)} y_i y_j K^{(-f)}(\mathbf{x}_i, \mathbf{x}_j) \right| \end{aligned} \tag{5}$$

where:

c_f means change of the objective function as a result of *f*-th feature removal,

$K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function,

$\alpha_i^*, \alpha_i^{*(-f)}$ are respective solution of the SVM equations [12].

2. Computer Methods and Theory

There are four classes of globular proteins defined as follows [13]:

α class is composed of mainly α helices with occurring connection between them,

β class is composed of β sheets with occurring connection between them,

α/β class contains α helices as well as β sheets occurring in an alternating manner with connection between them and to a large measure with parallel β sheets,

$\alpha+\beta$ class with separated areas both: α and β areas. Antiparallel β sheet structures occur.

Our investigation deals with feature selection of the structural classes by the aid of the SVM classifier. The great advantage of this linear and binary classifier is setting of an optimal separating hyper-plane which minimizes generalization error [14÷16].

2.1. Data Set

SCOP approach based on sequence identity and structure similarity seems to be the most reliable and comprehensive. The SCOP database is organized in several levels of so called evolutionary hierarchy with the main structural classes on the top [17, 18]. The domain as a basic classification entity was used as proposed by Murzin from the SCOP database based on structural and sequential similarity and so called evolutionary relationship [19].

The data were split into three data pools: training, test and validation. The classic 30% of paired identity threshold of significant homology was applied to avoid data redundancy and compare to the other application [20]. The composition of amino acids (AAC) and pseudo composition (PseAA) are applied as features of classification tests [21–29].

2.2. Feature Selection

The RFE method avails the mechanism of the binary SVM classifier, in which the criterion for assessing the weights of features was used in the objective function [30]. In this paper the multi-class selection was determined by scaling the relative distance between the m features for each n binary classification. After averaging these values for all n binary classifiers, weights were obtained for each m features, which were the basis for determining the ranking for multi-class classification:

$$c_f = \frac{1}{n}\sum_{k=1}^{n}\frac{\left\|w_k^{(-c\max)}\right\|^2 - \left\|w_k^{(-f)}\right\|^2}{\left\|w_k^{(-c\max)}\right\|^2 - \left\|w_k^{(-c\min)}\right\|^2} \tag{6}$$

hence:

$$c_f = \frac{1}{n}\sum_{k=1}^{n}\frac{\sum_{i,j=1}^{M}\alpha_{ik}^{*}\alpha_{jk}^{*}y_{ik}y_{jk}K_k^{(-c\max)}(\mathbf{x}_{ik},\mathbf{x}_{jk}) - \sum_{i,j=1}^{M}\alpha_{ik}^{*}\alpha_{jk}^{*}y_{ik}y_{jk}K_k^{(-f)}(\mathbf{x}_{ik},\mathbf{x}_{jk})}{\sum_{i,j=1}^{M}\alpha_{ik}^{*}\alpha_{jk}^{*}y_{ik}y_{jk}K_k^{(-c\max)}(\mathbf{x}_{ik},\mathbf{x}_{jk}) - \sum_{i,j=1}^{M}\alpha_{ik}^{*}\alpha_{jk}^{*}y_{ik}y_{jk}K_k^{(-c\min)}(\mathbf{x}_{ik},\mathbf{x}_{jk})} \quad (7)$$

While Cmax and Cmin are the numbers of features for which the function

$$\sum_{i,j=1}^{M}\alpha_{ik}^{*}\alpha_{jk}^{*}y_{ik}y_{jk}K_k^{(-c)}(\mathbf{x}_{ik},\mathbf{x}_{jk}) \quad (8)$$

takes the value of the maximum and minimum, respectively ($c = 1, ..., m$).
$\mathbf{x}_{ik}$ – is the i-th support vector of k-th classifier, $i = 1, ..., s_k$, where s_k is the amount of support vectors of k-th binary classifier.

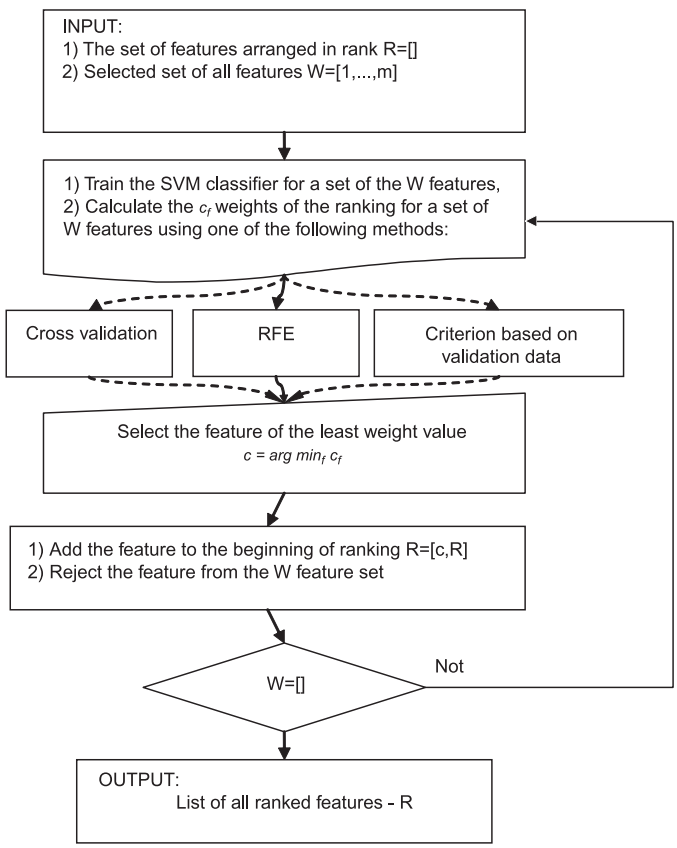

Fig. 1. Algorithm of feature selection. R – set of ranked features, W – set of unselected features, c – number of the least relevant feature

In this study, three methods based on the RFE schema were applied (see flow chart, Fig. 1). Only the way of determining the weights associated with the used method is variable. The criterion of cross-validation method was CV_f coefficient after *f*-th feature removal, where $c_f = 1 - CV_f$.

For the RFE method, the weight value is determined by the formula 7. In the third method, the weights are determined by a set of validation data for which the classification error is determined after the removal of the considered feature.

In order to eliminate features that do not substantially affect or have a negative impact on accuracy of the classification, selection of the features is performed. With reduction of the features, both reducing the amount of stored data dimensions and improvement of processing performance as well as a better understanding of the impact of the features on the classification accuracy are achieved.

3. Results and Discussion

3.1. Feature Construction

Following four types of the feature sets were taken into consideration: standard amino acid composition (AAC), pseudo amino acid composition with various variants and *1/n* feature reflecting the length of protein domain sequence (tests, model selection and construction of features is described with details in [21]).

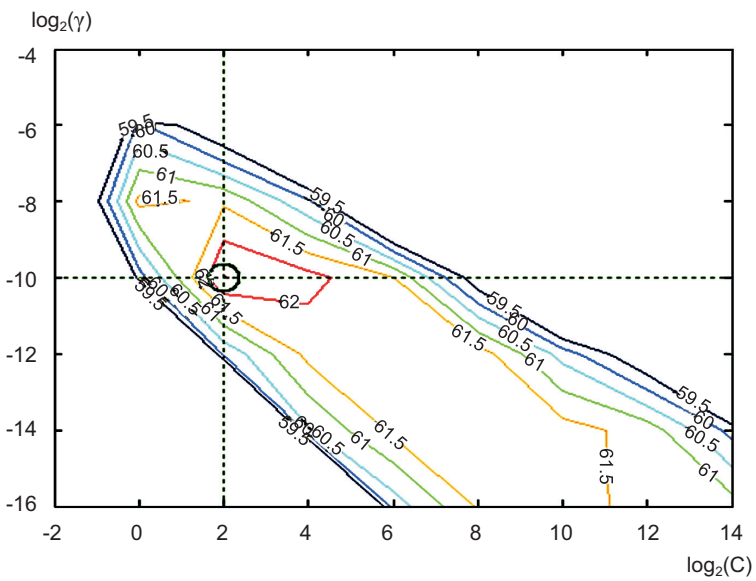

Fig. 2. Contour diagram obtained by the grid, for the SVM classifier using the features of PSE Type3+*1/n.*]. γ, c – two control parameters of the SVM classifier for the RBF kernel

The best classification accuracy was obtained for the following feature construction:

$$\begin{cases} \tau_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} J_{i,i+1}^{1} \\ \tau_2 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} J_{i,i+1}^{2} \\ \vdots \\ \tau_k = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} J_{i,i+1}^{k} \\ \tau_{k+1} = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2} J_{i,i+2}^{1} \\ \vdots \\ \tau_{2k} = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2} J_{i,i+2}^{k} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (\lambda < L) \\ \tau_{k\lambda-1} = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda} J_{i,i+\lambda}^{1} \\ \tau_{k\lambda} = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda} J_{i,i+\lambda}^{k} \end{cases} \tag{9}$$

where: $J_{i,j}^{m} = [h_m(R_i) - h_m(R_j)]^2$.
The features are constructed as follows:

$$P = \begin{bmatrix} p_1 \\ \vdots \\ p_{21} \\ p_{21+1} \\ \vdots \\ p_{21+k\lambda} \end{bmatrix} \tag{10}$$

and

$$P_u = \begin{cases} \dfrac{f_u}{\displaystyle\sum_{i=1}^{21} f_i + w\sum_{j=1}^{k\lambda} \tau_j}, & 1 \le u \le 21 \\ \dfrac{w\tau_{u-21}}{\displaystyle\sum_{i=1}^{21} f_i + w\sum_{j=1}^{k\lambda} \tau_j}, & 21+1 \le u \le 21+k\lambda \end{cases}. \tag{11}$$

Feature *1/n* is added as a feature representing the length, where *n* is the domain length and could be treated as a twenty first amino acid with appearance frequency inversely proportional to protein domain length. The results were achieved for the values: $w = 1$ and $\lambda = 15$ with the number of biochemical properties $k = 6$. Biochemical values are as follows: hydrophobicity, hydrophilicity, mass of side chain, pK1, pK2 and pI. For these parameters, the accuracy rate was achieved for the method of grid-CV equal to 62.52% for 111 features and the coefficients $C = 2^2$ and $\gamma = 2^{-10}$ (Fig. 2).

3.2. Feature Selection

According to the assumptions three following methods were applied to reduce amount of the features:

1) Nonlinear multi-class RFE method,
2) Cross Validation method (CV),
3) Reduction by means of separate validation data.

We performed the feature selection for the features with the best accuracy rate. We adopted CV accuracy rate for training data as a basic criterion of the feature selection method evaluation. The best results were obtained using the CV method. The best value of CV was obtained for 72 features, and it stood at 63.87% and classification accuracy for the test data of 62.07% (Table 1).

Table 1. Accuracy of classification for groups of selected features: all 111 features, 72 features of the best CV ratio, 30 features of CV ratio as before selection. Amount of features – amount of selected features CV for training data – cross validation coefficient for selected features

Amount of features	CV (%) for training data	Classification accuracy for test data (%)
111	63	62
72	64	62
30	63	62

Assuming a maximum reduction of the features for the CV accuracy rate not less than before the reduction, one can make the elimination of up to 30 features. The CV accuracy rate is 62.57% while the classification accuracy of 61.79% for the test data.

The data set chosen in this way could provide a set of features which is a base for further investigation of the influence of particular amino acids interactions on classification with respect to their biochemical properties and offset coefficients for each correlation.

Only thirty the most important features were selected from among one hundred eleven features without loss of the classification quality.

Figure 3 shows that during the reduction of most features, the classification quality remains at a similar level (Met. CV). Significant decrease is noticed if the reduction is continued for the last 30 selected features. This means that the relevance and thus the credibility of each feature, is greater in a selected group than the relevance of the eliminated features in their groups.

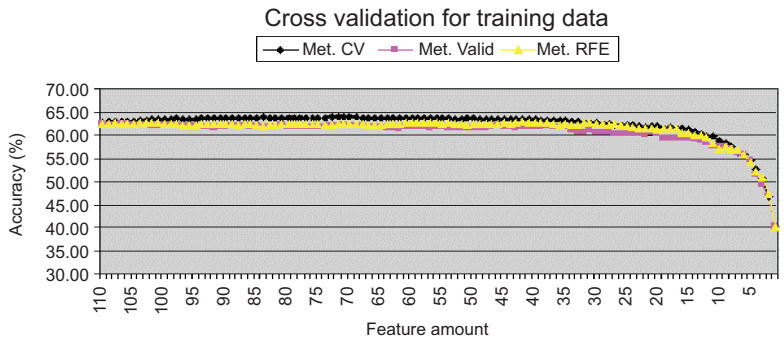

Fig. 3. Effect of the feature selection on cross-validation rate for the training data and the features of PSE – Type 3. Methods: CV, validation, RFE. Accuracy – cross validation coefficient for three mentioned feature selection methods. Feature amount – amount of the relevant features

3.3. PCA Dimensionality Reduction

Figure 4 shows the descending order of eigenvalues associated with the new features defined by projection onto the corresponding eigenvectors. The first five features have by far the largest variance. However, a visual or mathematical formula to eliminate the features with the lowest variance will not guarantee an optimal reduction of the new features to offer the best classification accuracy. In the present application, dimension reduction was made in groups of ten features with the lowest variance. The classifier was used as a tool for verifying the validity of the elimination of the individual groups of features. Because the ranking does not depend on the classifier used for each of the tested groups of features, a new model with new parameters was being set in order to increase its effectiveness. The CV coefficient was calculated again for the training data and accuracy for the test data (the test data consisted of the previous test and validation data; here the split over the test data and the validation data was not applied).

The CV coefficients of the cross validation method of feature selection from the start values are above 62% even though the test for a set amount to 62% until the selection of 21 features. For 11 features, this value is 59% (Table 4 in appendix).

In the case of PCA, the CV reduction coefficient of 63% and an accuracy of 62% are observed for a set of 41 features with the largest variance. The CV-value and accuracy of the test data for the PCA method are already 60% and 58% for 31 features and respectively only 57% and 55% for 11 features (Table 2).

The PCA method is probably the most widely used method of reducing the dimension which forms entirely new uncorrelated features as a linear combination of original features.

But the problem with this type of reduction for classification is that the PCA method does not include information on how the features of the examples belonging to each class. New features of low variance often have an important impact on the quality of the classification while the features of high variance may be irrelevant [31].

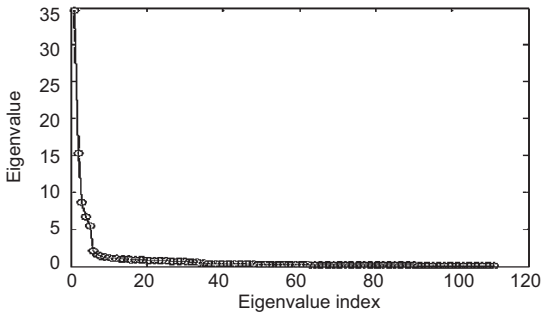

Fig. 4. Descending order of eigenvalues (vertical axis) associated with the new features defined by projection onto the corresponding eigenvectors (Eigenvector index)

Table 2. Classification accuracy of the reduction results of sample every 10 features with the lowest variance. CV – the cross validation ratio for the training data. Accuracy valid+test – classification accuracy for the combined test data and validation. RBF g, c – best parameters g and C determined using the grid method for the SVM classifier

Amount of features	CV	Accuracy valid+test	RBF g;c
111	61	60	–24;18
101	61	61	–22;14
91	61	60	–18;14
81	62	61	–18;16
71	62	61	–18;16
61	62	61	–10;4
51	62	61	–18;14
41	63	62	–14;12
31	62	61	–10;4
21	60	58	–14;14
11	57	55	–8;10

3.4. Discussion on Selected Features

Rank of the relevant features is presented in Table 3. The name of amino acids is specified for the features related to amino acid composition. Biochemical property and offset of *u*-tier correlation factor are related to the PseAA features. The set of 30 most relevant features is composed of 10 normalized AAC features, 1/*n* feature and 19 features based on correlation of the biochemical properties.

One of the key elements allows the prediction of α helices as well as *β* sheets is the propensity of amino acids to form or break these structures [32–34]. In Table 1 of [32], the following groups of propensity to form or brake second order structures were determined for α helices and β sheets respectively: H_α, H_β – strong formers, h_α, h_β – formers, I_α, I_β – week formers, i_α, i_β – indifferent, b_α, b_β – breakers, B_α, B_β – strong breakers.

Table 3. A detailed description of the selected 30 most relevant features using the CV method. The first column specifies the rank of the most relevant 30 features. The feature number is the number assigned to each feature as follows : {1-Ala,2-Cys,3-Asp,4-Glu,5-Phe,6-Gly,7-His,8-Ile,9-Lys,10-Leu,11-Met,12--Asn,13-Pro,14-Gln,15-Arg,16-Ser,17-Thr,18-Val,19-Trp,20-Tyr,21-1/n} and formula 9 for $\lambda = 15$ [21]

Ranking	Feature number	Rest	Property	u
30	40	----------	Hydrophobicity	4
29	73	----------	pK1	9
28	51	----------	pI	5
27	52	----------	Hydrophobicity	6
26	96	----------	Mass	13
25	45	----------	pI	4
24	61	----------	pK1	7
23	99	----------	pI	13
22	76	----------	Hydrophobicity	10
21	22	----------	Hydrophobicity	1
20	43	----------	pK1	4
19	19	Tryptophan	----------	---
18	59	----------	Hydrophilicity	7
17	95	----------	Hydrophilicity	13
16	4	Glutamic acid	----------	---
15	17	Threonine	----------	---
14	35	----------	Hydrophilicity	3
13	11	Methionine	----------	---
12	10	Leucine	----------	---
11	13	Proline	----------	---
10	8	Isoleucine	----------	---
9	30	----------	Mass	2
8	34	----------	Hydrophobicity	3
7	41	----------	Hydrophilicity	4
6	28	----------	Hydrophobicity	2
5	6	Glycine	----------	---
4	1	Alanine	----------	---
3	18	Valine	----------	---
2	29	----------	Hydrophilicity	2
1	21	1/n	----------	---

So, it seems that the features related to occurrence of strong formers as well as strong breakers should play the key role for diversity of structural classes which are closely connected with secondary structures. These amino acids well determine the presence as well as absence of the regions of regular secondary structures. Glutamic acid (Glu), alanine (Ala) and leucine (Leu) belong to α strong former group H_α. β strong former group H_β consist of methionine (Met), valine (Val) and isoleucine (Ile). Amino acids which strongly break of α helix of group B_α are proline (Pro) and glycine (Gly) while Glutamic acid (Glu) belongs to β strong breaker group B_β. There

are all eight amino acids of H_α, H_β and B_α, B_β groups among ten amino acids representing the AAC features of 30 the most relevant features. Unclear, however, may be the role of the features associated with the occurrence of threonine and tryptophan. Tryptophan is amino acid belonging to the group forming the structure of both β sheet and α helices. Threonine is amino acid of the group that forms β sheet structure and group indifferent for the formation of α helix. It is worth noting that tryptophan is the most compact and has the largest mass of all twenty amino acids. Other features are related to the interaction between neighboring amino acid residues, and surface of structures formed by them, on the basis of their biochemical properties. You can observe that the features connected with correlation of the biochemical properties have the relevant influence for every second, every third and every forth rests of the domain sequence. It seems that the mentioned features should be associated with influence of interaction between neighboring amino acid residues on stability protein structures of β sheet for every second amino acid rest and in case of α helix structure for every third and every forth amino acid rest. Every second rest of the β sheet forms a surface which interacts with the surface of an adjacent structure or solvent [35, 36]. Every second rests of the β sheet and every third and forth rests of the α helix are located in close proximity to each other and can form local interactions [37–41]. Also, every third or every fourth side chain of the α helix form characteristic edges involved in the interactions between other helical structures [42, 43]. Every forth rest of the α helix of two adjacent rows play key role in packing of the α helix and the β sheet [44–49].

4. Conclusion

The best results of classification were achieved for the features reflecting the order effect, based on the biochemical properties together with the *1/n* feature. The feature selection was performed according to the RFE main algorithm on the basis of three criteria: the CV coefficient, the classification accuracy of test data and the SVM objective function increase (RFE). The CV coefficient was chosen for assessing a specific pool of features. Although the feature selection method based on the CV coefficient is computationally the most time-consuming method, it allows the selection of the best set of features (due to the criterion of maximizing the coefficient of CV). The feature selection made amount reduction of the used features to a set of 30 traits possible without loss of the SVM classification quality. Selection of the relevant features made reduction of a larger number of features possible while maintaining the quality level of classification than the reduction of features by using the PCA method. Unlike the PCA, the feature selection method allows the separation of the essential original features, which are the basis of further analysis and interpretation of the specific amino acid influence together with features which determine the specific biochemical properties in a certain order of protein domain sequence.

Table 4. The results of feature selection following methods: CV, Valid and RFE for PseAA type 3 features. Every tenth sample features were presented. The first column (*CV*) specifies the cross validation ratio for the training data. The second (*accuracy valid*), third (*accuracy test)* and fourth (*accuracy valid+test*) columns include the accuracy ratios of classification for the data, respectively, validation, testing, testing & validation together. The last column (*Feature number*) is the ranking of features for the specified methods

	CV (%)			**Accuracy (%) valid**			**Accuracy (%) test**			**Accuracy (%) valid+test**			**Feature number**		
No	CV	Valid	RFE	CV	Valid	RFE	CV	Valid	RFE	CV	Valid	RFE	CV	Valid	RFE
110	62.78	62.42	62.65	61.67	61.98	61.46	62.47	62.73	62.37	62.07	62.36	61.92	91	90	109
100	63.33	62.16	62.52	61.51	62.97	61.61	62.58	61.75	62.32	62.05	62.36	61.97	66	53	49
90	63.66	61.74	62.62	61.41	63.91	62.29	62.63	62.06	62.47	62.02	62.98	62.38	31	55	98
80	63.56	61.76	62.81	60.99	63.49	61.20	62.11	61.96	62.68	61.55	62.72	61.94	69	89	48
70	63.79	61.87	62.60	60.94	63.75	61.67	63.20	62.32	62.42	62.07	63.03	62.05	74	64	93
60	63.56	61.69	62.70	60.99	63.18	61.56	62.37	61.96	62.58	61.68	62.56	62.07	93	63	73
50	63.61	61.48	62.47	61.04	63.28	61.67	62.89	61.70	62.37	61.97	62.49	62.02	92	111	63
40	63.33	61.63	62.42	60.78	62.76	61.56	62.47	61.75	62.63	61.63	62.25	62.10	42	92	96
30	62.57	60.54	61.53	60.73	62.45	60.83	62.84	61.91	62.99	61.79	62.18	61.92	88	7	4
20	61.84	60.23	61.30	60.99	61.82	60.57	62.27	61.34	61.55	61.63	61.58	61.06	22	22	22
10	58.75	57.52	58.54	58.33	60.68	60.10	59.85	58.56	59.48	59.09	59.61	59.79	13	58	52
1	40.27	40.27	40.27	40.83	40.83	40.83	40.62	40.62	40.62	40.73	40.73	40.73	29	29	28
													21	21	21

References

1. Guyon I., Weston J., Barnhill S., Vapnik V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 2002, 46, 389–422.
2. Guyon I., Vapnik V., Boser B., Bottou L.: Structural Risk Minimization for Character Recognition. S.A. Solla, AT&T Bell Laboratories, Holmdel, USA 1992.
3. Dutkowski J.: Exploratory data analysis. Projection methods: principal component analysis and multidimensional scaling. (in Polish). http://www.mimuw.edu.pl/~aniag/SADM/pca.pdf (last access 08.08.2012).
4. Twardowski T.: Numerical methods for technical computing, Lecture VII, Eigenvalues and eigenvectors, singular values and SVD decomposition (in Polish). http://galaxy.uci.agh.edu.pl/~ttward/numer/Warto%9Cci%20i%20wektory%20w%B3, asne.pdf (last access 08.08.2012).
5. Kohavi R., John G.: Wrappers for Feature Subset Selection. Artificial Intelligence, December 1997, 97, 1–2, 273–324.
6. Guyon I., Elisseeff A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 2003, 3, 1157–1182.
7. Liu J., Ranka S., Kahveci T.: Classification and feature selection algorithms for multi-class CGH data. Bioinformatics 2008, July, 24, 13.
8. Guyon I.: Feature selection and causal discovery fundamentals and applications. http://langtech.jrc.it/mmdss2007/htdocs/Presentations/Docs/MMDSS_Guyon.pdf (last access 08.08.2012).
9. Guyon I., Elisseeff A.: An Introduction to Feature Extraction, Feature Extraction. Foundations and Applications. Springer 2006.
10. Le Cun Y., Denker J., Solla S.: Optimal Brain Damage. AT&T Bell Laboratories, Holmdel, N. Y. 1990.
11. Zhou X., Tuck D.: MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Department of Pathology, Yale University School of Medicine, New Haven, Connecticut 2007.
12. Abe S.: Support Vector Machines for Pattern Classification. Springer 2005.
13. Levitt M., Chothia C.: Structural patterns in globular proteins. Nature 1976, June 17, 261, 5561, 552–558.
14. Osuna E., Freund R. Girosi F.: A.I. Memo : Support Vector Machines Training and Applications No. 1602, C.B.C.L Paper No. 144, 1997.
15. Vapnik V.: The Nature of Statistical Learning Theory. Second Edition, Springer 1995.
16. Fradkin D., Muchnik I.: Support Vector Machines for Classification. IMACS Series in Discrete Mathematics and Theoretical Computer Science 2005.
17. Hubbard T., Ailey B., Brenner S., Murzin A., Chothia C.: SCOP, Structural Classification of Proteins Database: Applications to Evaluation of the Effectiveness of Sequence Alignment Methods and Statistics of Protein Structural Data. Acta Cryst. 1998, D54, 1147–1154.
18. Hubbard T., Ailey B., Brenner S., Murzin A., Chothia C.: SCOP, Structural Classification of Proteins Database. Nucleid Acids Research 1999, 27, 1.
19. Murzin A., Brenner S., Hubbard T., Chothia C.: SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. J. Mol. Biol. 1995, 247, 536–540.
20. Gu F., Chen H., Ni J.: Protein structural class prediction based on an improved statistical strategy. BMC Bioinformatics 2008, 9 (Suppl 6):S5.
21. Krajewski Z.: Protein structural classification based on pseudo amino acid composition using SVM classifier. Biocybernetics and Biomedical Engineering 2013, vol 33 (in print).
22. Chou K.: Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. Proteins: Structure, Function, and Bioinformatics, 2001, 43, 3, 246–255.
23. Chou K., Cai Y.: Predicting Protein Quaternary Structure by Pseudo Amino Acid Composition. Proteins: Structure, Function, and Genetics 2003, 53, 282–289.

24. Zhang G., Li H., Gao J., Fang B.: Predicting Lipase Types by Improved Chou's Pseudo-Amino Acid Composition. Protein and Peptide Letters 2008, 15, 10, 1132–1137.
25. Chou K.: Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. Current Proteomics, 2009, 6, 262–274.
26. Chou K., Cai Y.: Prediction of protease types in a hybridization space. Biochemical and Biophysical Research Communications 2006, 339, 1015–1020.
27. Chou K., Cai Y.: Predicting Subcellular Localization of Proteins by Hybridizing Functional Domain Composition and Pseudo-Amino Acid Composition. Journal of Cellular Biochemistry 2004, 91, 6, 1197–1203.
28. Chou K.: Progress in Protein Structural Class Prediction and its Impact to Bioinformatics and Proteomics. Current Protein and Peptide Science October 2005, 6, 5, 423–436.
29. Cai Y., Zhou G., Chou K.: Support Vector Machines for Predicting Membrane Protein Types by Using Functional Domain Composition. Biophysical Journal, 1 May 2003, 84, 3257–3263.
30. Shieh M., Yang C.: Multiclass SVM-RFE for product form feature selection. Expert Systems with Applications, July-August 2008, 35, 1–2, 531–541.
31. Oza N., Turner K.: Dimensionality Reduction Through Classifier Ensembles. Technical Report NASA-ARC-IC-1999-126, NASA Ames Research Center, 1999.
32. Fasman G. (Editor) : Prediction of Protein Structure and the Principles of Protein Conformation. Springer; 1 ed. (October 31, 1989). Chapter 9: Prevelige P., Fasman G.: Chou-Fasman Prediction of the Secondary Structure of Proteins The Chou-Fasman-revelige Algorithm.
33. Edholm O.: The Chou-Fasman method for predicting secondary structure. Alba Nova University Center, KTH - Theoretica Physics , SE-106 91 Stockholm – Sweden.
34. Singh M.: COS551 Intro. to Computational Biology. http://www.cs.princeton.edu/~mona/Lecture/sec--structure.pdf (last access 08.08.2012)
35. Chothia C., Hubard T., Brenner S., Barns H., Murzin A.: Protein Folds in the all-α and all-β classes. Annu. Rev. Biophys. Biomol. Struct. 1997, 26, 597–627.
36. Murzin A., Lesk A., Chothia C.: Principles Determining the Structure of β-Sheet Barrels in Proteins. J. Mol. Biol. 1994, 236, 1369–1381.
37. Muńoz V., Cronet P., López-Hernández E., Serrano L.: Analysis of the effect of local interactions on protein stability. Folding and Design, June 1996, 1, 3, 167–178.
38. Muñoz V., Serrano L.: Local versus nonlocal interactions in protein folding and stability – an experimentalist's point of view. Folding and Design, August 1996, 1, 4, R71–R77.
39. The Biochemistry Questions. http://biochemistryquestions.wordpress.com/2008/10/02/secondary--structure-of-proteins/ (last access 08.08.2012).
40. Alpha-Helix: Overview of Secondary Structure. http://mcdb-webarchive.mcdb.ucsb.edu/sears/biochemistry/ (last access 08.08.2012).
41. Overview of Beta-Pleated Sheet Secondary Structure. http://mcdb-webarchive.mcdb.ucsb.edu/sears/biochemistry/ (last access 08.08.2012).
42. Chothia C.: Polyhedra for helical proteins. Nature, 19 January 1989, 337.
43. Chothia C.: Principles that determine the structure of proteins. Ann. Rev. Biochem. 1984, 53, 537–72.
44. Chothia C., Levitt M., Whaildson D.: Helix to Helix Packing in Proteins. J. Mol. Biol. 1981, 145, 215–250.
45. Chothia C., Levitt M., Richardson D.: Structure of proteins: Packing of a-helices and pleated sheets. Proc. Nati. Acad. Sci. USA October 1977, 74, 10, 4130–4134.
46. Chou K., Carlacci L.: Energetic Approach to the Folding of α/β Barrels. Proteins: Structure, Function and Genetics, 1991, 9, 280–295.
47. Chothia C., Finkelstein A.: The classification and origins of protein folding patterns. Annu. Rev. Biochem. 1990. 59:1007–39.

48. Janin J., Chothia C.: Packing of α-Helices onto β-Pleated Sheets and the Anatomy of α/β Proteins. J. Mol. Biol. 1980, 143, 95–128.
49. Chothia C., Janin J.: Orthogonal Packing of β-Pleated Sheets in Proteins. Biochemistry 1982, 21, 3955–3965.