The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

SVM-based Method for Predicting Enzyme Function in a Hierarchical Context

Yong-Cui Wang[1] Zhi-Xia Yang[2] Nai-Yang Deng[1,*]

[1]College of Science, China Agricultural University, Beijing 100083, China
[2]College of Mathematics and System Science, Xinjiang University, Urumuchi 830046, China

Abstract Automatically categorizing enzyme into the Enzyme Commission (EC) hierarchy is crucial to understand its specific molecular mechanism. Standard machine learning methods like support vector machine (SVM) and naïve bayesian classifier have been successfully applied for this task. However, they treat each functional class independently, and ignore the inter-class relationships. In this paper, we develop a SVM-based method for prediction of enzyme function into the EC hierarchical context. Our method with low computational complexity is a modified version of a structured predictive model—Hierarchical Max-Margin Markov algorithm (HM^3). HM^3, which is specially designed for the hierarchical multi-label classification, has been successfully used in many structured pattern recognition problems, such as document categorization, web contend classification, and enzyme function prediction. As input features for our predictive model, we use the conjoint triad feature (CTF). Our method has been validated on an enzyme benchmark dataset, the proteins in this benchmark dataset have less than 40% sequence identity to any other in a same functional class. Finally, for the first three EC digits, the predictive accuracy and the Matthew's correlation coefficient (MCC) of our method range from 78% to 100% and 0.76 to 1 respectively. Therefore we think our new method will be useful supplementary tools for the future studies in enzyme function prediction.
Keywords: Enzyme function prediction; Conjoint triad feature; Structured hierarchical output; Support vector machine.

1 Introduction

About half of all the proteins have been characterized as function of enzymatic activity by various biochemical experiments. In addition, accurate assignment of enzyme function is a prerequisite of high-quality metabolic reconstruction and the analysis of metabolic fluxes [1]. It should be noted that, the Enzyme Commission (EC) number is comprised of four digits separated by periods to classify the functions of enzymes [2]. The first three digits describe the overall type of an enzymatic reaction, while the last digit represents the substrate specificity of the catalyzed reaction. Since the first three digits have the direct relationship with enzyme function, it is in pressing need to develop computational methods to accurately output the first three EC digits for a given protein.

*Corresponding author. E-mail: dengnaiyang@cau.edu.cn

The straightforward idea is to transfer an EC number between two globally aligned protein sequences. The accuracy of the direct inference has been reported to significantly drop under 60% sequence identity [3]. One of the remedies is to develop computational methods to automatically categorize enzyme into EC hierarchy. Some works concentrate in predicting the top lever of the EC taxonomy, and then outputting the second and third digits of EC number. For example, Chou et.al [4] developed a top-down approach for predicting enzyme functional classes and sub-classes, and the overall accuracy for the first two layers is higher than 90%. However, they treated functional class independently, and ignore the inter-class relationships. To address this limitation, some works aim at the structured prediction of enzyme functions in the whole EC hierarchical taxonomy. For example, Astikainen et.al. [5] introduced the Hierarchical Max-Margin Markov (HM^3) to output the hierarchical enzyme function, and obtained over 85% accuracy on the four digits EC number. HM^3, which is specially designed for the hierarchical multi-label classification, generalizes support vector machine (SVM) learning and that is based on discriminant functions that are structured in a way that mirrors the class hierarchy [7]. It has been successfully used in many structured pattern recognition problems, such as document categorization [7], web contend classification [8] and so on. However, the HM^3 is time consuming, and unsuitable for the large predictive task. For dealing with the computational challenge of the HM^3, in this article, we reformulate the model of HM^3, and make the number of the variable decrease to the number of the training samples.

Another key problem in predicting enzyme function is to encode a protein as a real-value vector. In previous studies, the amino acid composition (AAC) representation has been widely utilized in predicting enzyme family and subfamily class [9]. Owing to lack of the sequence order information, some modified versions of AAC, such as pseudo amino acid composition (Pse-AAC) [10] and amphiphilic pseudo-amino acid composition (Am-Pse-AAC) [11] have been developed. However, both Pse-AAC and Am-Pse-AAC have some parameters to be determined, and need the properties of physio-chemistry of amino acids. Recently, a much simple feature encoding method for protein-protein interactions (PPIs) prediction was proposed [12]. The authors have shown that SVM with the conjoint triad feature (CTF) outperformed other sequence-based PPI prediction methods. The CTF considers not only properties of one amino acid but also its vicinal amino acids and treats any three continuous amino acids as an unit. That is, it contains not only the composition of amino acids but also sequence-order effect. It has also successfully been used in prediction of DNA- and RNA-binding proteins [13]. Inspired by these, in this paper, we introduce the CTF into our structured predictive model, and the better results are expected.

The paper is structured as follows. In Materials and Methods section, we give the depiction for the CTF and benchmark dataset, and discuss the procedure for formulating the structure-based predictive model In Results section, we compare our method with the existing methods. Lastly, we conclude this paper.

2 Materials and Methods

2.1 Materials

The benchmark dataset used to validate the performance of our method is collected from the literature [4]. The sequences in this dataset have less than 40% sequence identity to any other in a same functional class. The detail information of this dataset can be

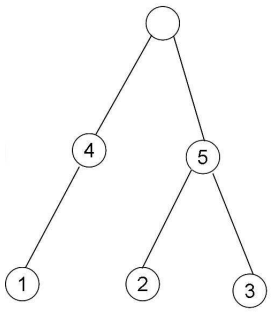

Figure 1: The toy example for functional hierarchical tree

found in [4]. In addition, for avoiding the extreme subfamily bias, those subfamilies which contain less than 40 proteins are excluded in our validation. Finally there are six main functional classes and thirty-four subfamily classes (i.e., twelve for oxidoreductases, seven for transferases, five for hydrolases, four for lyases, four for isomerases and two for ligases) in the benchmark dataset.

2.2 Methods

2.2.1 Input feature: CTF

Now let us construct the CTF. Based on the dipoles and volumes of the side chains, the 20 amino acids can be classified into seven classes: $\{A,G,V\}$, $\{I,L,F,P\}$, $\{Y,M,T,S\}$, $\{H,N,Q,W\}$, $\{R,K\}$, $\{D,E\}$, $\{C\}$. Thus, a $(7\times 7\times 7 =)343$-dimension vector is used to represent a given protein, where each element of this vector is the frequency of the corresponding conjoint triad appearing in the protein sequence. The detail description for the CTF can be found in [12].

2.2.2 Structured SVM

Now we give the detail representation for our structured predictive model. And we call it as the structured SVM.

Given examples $\{x_i, y_i\}$ for $i = 1,\cdots,l$, where x_i is a vector in the input space R^n and y_i denotes the corresponding class category taking a value in the output space $\{1,\cdots,q\}$, where q is the total number of categories. In addition, the functional hierarchical tree, representing the relationships among the categories, is also known, where the number of leaf nodes in the tree is q . A tree with $q = 3$ is shown in Figure 1.

Suppose that, there are a total of s nodes in the hierarchical functional tree (In Figure 1, $s = 5$). To predict enzyme functions in the whole EC hierarchical taxonomy, we introduce a structured predictive model—HM^3 [6] firstly, and then modified it to reduce its computational complexity.

Introduce $\Lambda(y)$ for output y by

$$\Lambda(y) = (\lambda_1(y),\cdots,\lambda_s(y))^{\mathrm{T}}, \tag{1}$$

where

$$\lambda_j(y) = \begin{cases} 1, & if\ j \in path(y); \\ 0, & otherwise, \end{cases} \tag{2}$$

where $path(y)$ is the category tags for the path from the root to the leaf node y in the functional hierarchical tree. For example, in Figure 1, $path(2) = \{2,5\}$.

The primal and Lagrange dual optimization problem for HM^3 are

$$\min_{\mathrm{w},b,\xi} \quad \frac{1}{2}\|\mathrm{w}\|^2 + C \cdot \sum_{i=1}^{l} \xi_i, \tag{3}$$

$$\text{s.t.} \quad (\mathrm{w} \cdot \delta\Phi_i(y)) \geqslant 1 - \xi_i\ ,\ i = 1, \cdots, l, y \neq y_i, \tag{4}$$

$$\xi_i \geqslant 0\ , \delta\Phi_i(y) = \Phi(x_i, y_i) - \Phi(x_i, y)\ ,\ i = 1, \cdots, l, \tag{5}$$

and

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\sum_{Y \neq Y_i, Y' \neq Y_j} \alpha_{iy}\alpha_{jy'}((\Lambda(y_i) - \Lambda(y)) \cdot (\Lambda(y_j) - \Lambda(y')))K(x_i, x_j) - \sum_{i=1}^{l}\sum_{y \neg y_i} \alpha_{iy}\ , \tag{6}$$

$$\text{s.t.} \quad \alpha_{iy} \geqslant 0,\ \ i = 1, \cdots, l\ , y \neq y_i, \tag{7}$$

$$\sum_{y \neq y_i} \alpha_{iy} \leqslant C,\ \ i = 1, \cdots, l\ , y \neq y_i, \tag{8}$$

respectively, where $\mathrm{w} = (\mathrm{w}_1^{\mathrm{T}}, \cdots, \mathrm{w}_s^{\mathrm{T}})^{\mathrm{T}}$, $\phi : R^n \to \mathscr{H}$ is a mapping from the input space R^n to a Hilbert space $\mathscr{H}$, and

$$\Phi(x_i, y_i) = \begin{pmatrix} \lambda_1(y_i) \cdot \phi(x_i) \\ \cdots \\ \lambda_s(y_i) \cdot \phi(x_i). \end{pmatrix}. \tag{9}$$

Noticing that the number of the variables equals $l \cdot (q-1)$, so the problem (6)$\sim$(8) will be quite large when l and q is large. However, if we can find a label $y_i^* \neq y_i$, such that

$$y_i^* = \arg_{y \neq y_i} \min\{(\mathrm{w} \cdot \Phi(x_i, y_i)) - (\mathrm{w} \cdot \Phi(x_i, y))\}, \tag{10}$$

the inequality (4) will become $(\mathrm{w} \cdot \delta\Phi_i(y_i^*)) \geqslant 1 - \xi_i\ ,\ i = 1, \cdots, l$. So the Lagrange dual problem can be reformulated as follows:

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j((\Lambda(y_i) - \Lambda(y_i^*)) \cdot (\Lambda(y_j) - \Lambda(y_j^*)))K(x_i, x_j) - \sum_{i=1}^{l} \alpha_i, \tag{11}$$

$$\text{s.t.} \quad \alpha_i \geqslant 0,\ \ i = 1, \cdots, l\ , \tag{12}$$

$$\alpha_i \leqslant C,\ \ i = 1, \cdots, l\ . \tag{13}$$

The number of the variables for the problem (11)$\sim$(13) decreases to l. Note that, this problem is reduced to the standard SVM model by replacing the kernel function with $((\Lambda(y_i) - \Lambda(y_i^*)) \cdot (\Lambda(y_j) - \Lambda(y_j^*)))K(x_i, x_j)$.

How to choose a suitable y_i^* for each input x_i? The following theorem may provide some information for this task.

Theorem 1. *Let k,l is the arbitrary two class labels in leaf nodes, x is a test datapoint, $\alpha_{iy}, i = 1,\cdots,l$, $y \neq y_i$ are the optimal solutions of the problem* (6)∼(8)*. Then for every pair of label tags k and l, we have that,*

$$|(\mathrm{w}\cdot\Phi(x,k)) - (\mathrm{w}\cdot\Phi(x,l))|^2 \leq \|\Lambda(k)-\Lambda(l)\|^2 \sum_{j=1}^{s}(\mathrm{w}_j\cdot\phi(x))^2, \tag{14}$$

where

$$(\mathrm{w}_j\cdot\phi(x)) = \sum_{i=1}^{l}\sum_{y\neq y_i}(\lambda_j(y_i)-\lambda_j(y))\alpha_{iy}K(x_i,x). \tag{15}$$

Proof: The conclusion can be proved by using Cauchy inequality. □

From the Theorem 1, we can see that the class labels which belong to the same subfamily of the test label have the lowest upper bound of $|(\mathrm{w}\cdot\Phi(x,y)) - (\mathrm{w}\cdot\Phi(x,y^*))|$, compared to the other label $y' \neq y$. So in this paper, y_i^* is set to be one of the class belonging to the same subfamily of y_i.

2.3 Implementing predictive model and evaluation criterions

By replacing the kernel function with $((\Lambda(y_i)-\Lambda(y_i^*))\cdot(\Lambda(y_j)-\Lambda(y_j^*)))K(x_i,x_j)$, the freely available software for SVM implementation: LIBSVM (v.2.88) [14], can be used for solving our structured SVM model. The RBF kernel function is used here, and the penalty parameter C and the RBF kernel parameter γ are optimized by grid search approach with 3-fold cross-validation.

To evaluate the performance of proposed methods, we run the 10 fold cross-validation procedure. Besides for the accuracy, the Matthew's correlation coefficient (MCC)[15], which allows us to overcome the shortcoming of accuracy on imbalanced data [16], is used to further evaluate the performance of our method.

3 Results

3.1 The results on the benchmark dataset

Here we would like to display the promising performance of the structured SVM by validating it on a benchmark dataset (see Materials and Methods section). For comparison, we also train OET-KNN and AMSVM on the same benchmark dataset. OET-KNN (Optimized evidence-theoretic k nearest neighbor) classifier was successfully used in top-down predicting enzyme family and subfamily class [4], while AMSVM (SVM with arithmetic mean offset) is a modified version of the standard SVM, which is specially designed for imbalance classification problem [17]. It should be noted that, predicting enzyme function is an imbalance multi-class classification problem due to the fact that the number of proteins in each enzyme family and subfamily makes a great difference. Both OET-KNN and AMSVM predict the top level of EC hierarchy firstly, and then deal with the second and third level. OET-KNN only reports the predictive accuracy on the first two EC digits, so the MCC and the corresponding results on the third EC digit for OET-KNN are not shown in the following subsection.

The accuracies and MCCs for OET-KNN, AMSVM and the structured SVM in prediction of enzyme family classes are listed in Table 1. From Table 1, we can see that

Table 1: The accuracy and MCC of the current methods on the top level class in the EC hierarchy

Family name	OET-KNN		AMSVM		Structured SVM	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
$EC1 : Oxidoreductases$	0.89	–	0.92	0.89	**0.94**	**0.91**
$EC2 : Transferases$	0.93	–	0.94	0.92	**0.96**	**0.93**
$EC3 : Hydrolases$	0.94	–	0.96	0.93	**0.97**	**0.95**
$EC4 : Lyases$	0.83	–	0.86	0.85	**0.88**	**0.86**
$EC5 : Isomerase$	0.81	–	0.85	0.84	**0.87**	**0.85**
$EC6 : Ligases$	0.95	–	0.96	0.89	**0.97**	**0.90**

AMSVM specially designed for imbalance problem performs well than OET-KNN, and the structured SVM outperforms AMSVM not only on the accuracy but also on the MCC for all six enzyme main families. These results suggest that the more properties of dataset itself are incorporated into the predictive model, the more better results can be expected.

We use box plots to exhibit the variability of predictive accuracy for comparable methods among enzyme subfamily classes and sub-subfamily classes (Figure 2,3). The box plots for the variability of MCC are shown in Figure 4. For the second EC digit (enzyme subfamily class), the accuracy of OET-KNN ranges from 50% to 100%, and AMSVM makes the accuracy range from 75% to 95%. Although the median of accuracy for OET-KNN is larger than that for AMSVM, the difference between them is about 3% and the variation range of AMSVM is less than that of OET-KNN. It implies that AMSVM is more stable than OET-KNN, and have competitive performance of OET-KNN. While the minimum value of accuracy for the structured SVM is about 80%, and the maximum value reaches 100%. Furthermore, the structured SVM outperforms AMSVM and OET-KNN not only on the range of accuracy waved but also on the median of accuracy. For the third EC digit (enzyme sub-subfamily class), except for EC6 sub-subfamily class, the structured SVM outperforms AMSVM and OET-KNN not only on the range of accuracy waved but also on the median of accuracy. For both the second and the third EC digits, although the length of the MCC box for the structured SVM seems longer than AMSVM, but the structured SVM has a much higher MCC than AMSVM. All these results imply that the performance of the structured SVM is superior to those of the existing methods, which do not take the hierarchical structure of the dataset into account.

3.2 Comparison with other structure-based methods

In [5], besides HM^3, the authors have been introduced another structure-based predictive model—Maximum Margin Regression algorithm (MMR) [18] to output enzyme function. The MMR generats one-class SVM to structured output domain. It has the same the number of variables as our structured SVM and can be implemented easily. On a gold standard enzyme dataset containing 3090 proteins, for the all individual EC digits, the MMR achieved the best accuracy, while HM^3 achieved the nearly same results, and the HM^3 obtains the best F1 scores and MMR comes close second. When predicting sub-subfamilies, the HM^3 obtained the over 79% F1 score , and obtained over 89%

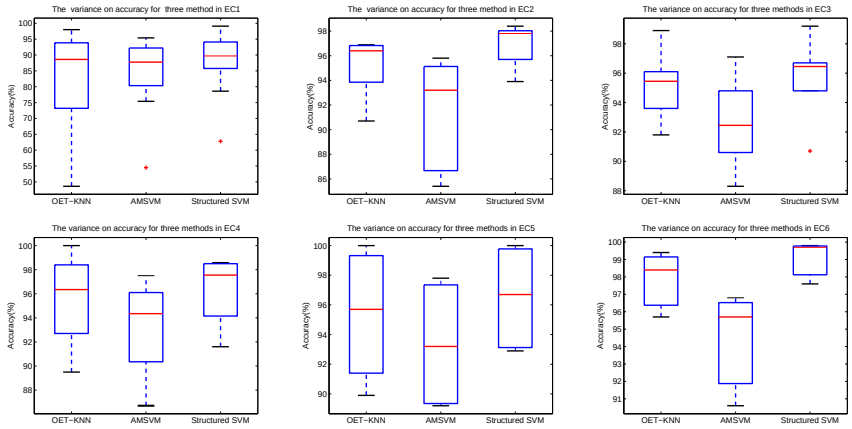

Figure 2: The box plots for accuracy on enzyme subfamily in EC1,EC2,EC3,EC4,EC5,EC6 respectively.

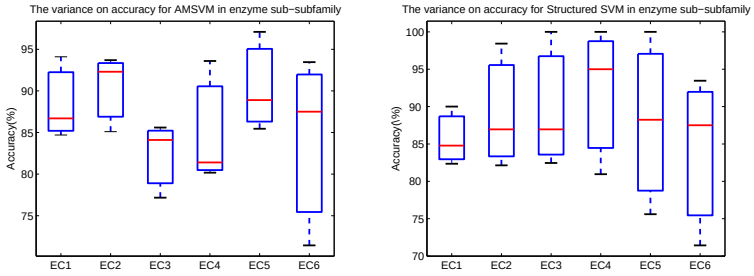

Figure 3: The box plots for accuracy on enzyme sub-subfamily.

F1 score when predicting the subfamilies and the main families [5]. Like the MCC, F1 score can avoid the bias due to the imbalance problem. In this work, our structured SVM has obtained the over 80% F1 score in predicting enzyme sub-subfamilies, and the over 90% F1 score in predicting subfamilies and the main families. These results suggest that our structured SVM has the competitive performance with the HM^3, and may outperform other existing low costly structure-based method, the MMR.

4 Conclusion

In this paper, we propose a structure-based method—the structured SVM for predicting enzyme function in EC hierarchy. The structured SVM is a modified version of HM^3 and can be easily implemented for the real-world large scale problem. The structured SVM has been validated on a benchmark enzyme dataset, where the best MCC and accuracy of 91% and 98% are obtained in predicting the main families. Furthermore, we obtain over 95% and 98% MCC in predicting subfamilies and sub-subfamilies respectively.

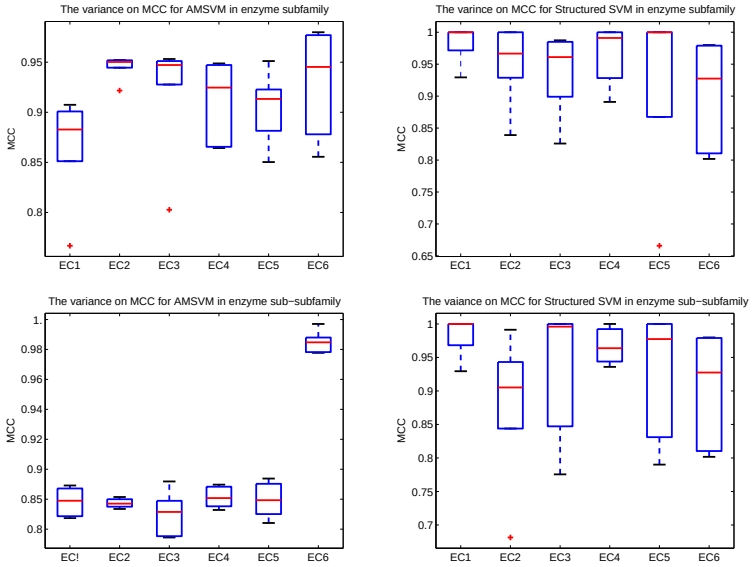

Figure 4: The box plots for MCC on enzyme subfamily (the top) and sub-subfamily (the down) respectively.

The predictive results suggest that for predicting the enzyme function in EC hierarchical taxonomy, the structured SVM performs much better than existing methods which do not take account of hierarchical relationship among enzyme categories. In addition, the structured SVM has exhibited the competitive performance with the HM^3, and the better predictive results than other existing low costly structured-based method (MMR) in structured output prediction of enzyme function. Therefore we think our new method will be useful supplementary tools for the future studies in enzyme function prediction.

Acknowledgments

This work is supported by the Key Project of the National Natural Science Foundation of China (No. 10631070), the National Natural Science Foundation of China (No. 10801112, No.10971223) and the Ph.D Graduate Start Research Foundation of Xinjiang University Funded Project (No.BS080101).

References

[1] Palsson, B. Systems Biology: Properties of Reconstructed Networks Cambridge University Press New York, NY, USA, 2006.

[2] Bairoch, A.The ENZYME database in 2000. *Nucleic Acids Research*, 2000, **28**, 304–305.

[3] Tian, W., Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, 2003, **333**, 863–882.

[4] Shen, H.B., Chou, K.C. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochemical and Biophysical Research Communications*, 2007, **364**, 53–59.

[5] Astikainen, K., Holm, L., Pitkänen, E., Szedmak, S., Rousu, J. Towards structured output prediction of enzyme function. *BMC Proceedings*, 2008, **2**:S2.

[6] Rousu, J., Saunders, C., Szedmak, S., Shawe-Taylor, J. Kernel-Based Learning of Hierarchical Multilabel Classification Models. *The Journal of Machine Learning Research*, 2006, *7*, 1601–1626.

[7] Cai, L, J., Hofmann, T. (2004) Hierarchical document categorization with support vector machines. Proceedings of the thirteenth ACM international conference on Information and knowledge management, Washington, D.C., USA.

[8] Dumais, S., Chen, H. Hierarchical Classification of Web Content, 2000, In SIGIR.

[9] Chou, K.C., Elrod, D.W. Prediction of enzyme family classes. *Journal of Proteome Research*, 2003, **2**, 183–190.

[10] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Genetics*, 2001, **43**, 246–255.

[11] Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 2005, **21**, 10–19.

[12] Shen, J.W., Zhang, J., Luo, X. M., Zhu, W.L., Yu, K.Q., Chen, K.X., Li, Y. X., Jiang, H.L. Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 2007, **104**, 4337–4341.

[13] Shao, X. J., Tian, Y. J., Wu, L. Y., Wang, Y., Deng, N. Y. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. *Journal of Theoretical Biology*, 2009, **258**, 289–293.

[14] Cheng, A.C., Coleman1, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C., Huang, E.S. Structure-based maximal affinity model predicts smallmolecule druggability. *Nature Biotechnology* , 2007, **25**, 71–75.

[15] Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 1975, **405**, 442–451.

[16] Pu, X., Guo, J., Leunga, H., Lin, Y.L.Prediction of membrane protein types from sequences and position-specific scoring matrices. *Journal of Theoretical Biology*, 2007, **247**, 259–265

[17] Li, B. Hu, J. Hirasawa, K. Sun, P., Marko, K. Support vector machine with fuzzy decision-making for real-world data classification. In IEEE World Congress on Computational Intelligence, Int. Joint Conf. on Neural Networks, 2006, Canada.

[18] Szedmak, S., Shawe-Taylor, J, Parado-Hernandez, E. Learning via Linear Operators: Maximum Margin Regression. Tech. rep., 2005, Pascal Research Reports.

[19] Wang, X.B., Wu, L.Y., Wang, Y.C., Deng, N.Y. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Engineering Design and Selection*, 2009, **22(11)**, 707–712.