Journal of Zhejiang University SCIENCE B
ISSN 1673-1581
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Science Letters:

EHPred: an SVM-based method for epoxide hydrolases recognition and classification*

JIA Jia (贾 佳)[1], YANG Liang (杨 亮)[1], ZHANG Zi-zhang (张子张)[†‡1,2]

(*[1]James. D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310008, China*)

(*[2]Department of Chemistry, Zhejiang University, Hangzhou 310027, China*)

[†]E-mail: zhangzz@zju.edu.cn

Received Aug. 12, 2005; revision accepted Oct. 23, 2005

Abstract: A two-layer method based on support vector machines (SVMs) has been developed to distinguish epoxide hydrolases (EHs) from other enzymes and to classify its subfamilies using its primary protein sequences. SVM classifiers were built using three different feature vectors extracted from the primary sequence of EHs: the amino acid composition (AAC), the dipeptide composition (DPC), and the pseudo-amino acid composition (PAAC). Validated by 5-fold cross tests, the first layer SVM classifier can differentiate EHs and non-EHs with an accuracy of 94.2% and has a Matthew's correlation coefficient (MCC) of 0.84. Using 2-fold cross validation, PAAC-based second layer SVM can further classify EH subfamilies with an overall accuracy of 90.7% and MCC of 0.87 as compared to AAC (80.0%) and DPC (84.9%). A program called EHPred has also been developed to assist readers to recognize EHs and to classify their subfamilies using primary protein sequences with greater accuracy.

Key words: Epoxide hydrolases (EHs), Amino acid composition (AAC), Dipeptide composition (DPC), Pseudo-amino acid composition (PAAC), Support vector machines (SVM)

doi:10.1631/jzus.2006.B0001 **Document code:** A **CLC number:** Q55

INTRODUCTION

Epoxide hydrolases (EHs) play a key role in the transformation and metabolism of xenobiotics in bio-systems (Armstrong, 1987). The enzymes are found ubiquitously in various organisms, including mammals, plants, insects and microorganisms, etc. (Argiriadi *et al.*, 1999). In the human body, EHs are found in organs such as the liver, lung, heart, brain, breast tissue, kidneys, and are associated with various diseases, including emphysema, pregnancy-induced hypertension, acute respiratory distress syndrome, inflammation and several forms of cancers (Fretland and Omiecinski, 2000). In most of these cases, the role of EHs, however, remains unknown. As a result, we are interested in investigating EHs (EC 3.3.2.3), particularly the relationships between its structures, its primary sequences, and its function features such as substrate specificity, selectivity, and kinetics, among others. However, the tertiary structures of EHs are mostly unsolved. Meanwhile, new sequence data accumulate exponentially as more than 400 EHs related sequences have been collected in the database (Barth *et al.*, 2004a), creating great demand for automatic methods of recognition and classification of known and newly sequenced EHs based on their primary sequences and structural features. Such tools would enable scientists to understand better the biochemical properties of EHs, and their associated human diseases. These potential new methods of EH analysis are of great value to many fields of research such as pharmaceuticals and the life sciences.

However, annotating and assigning the functions and the classes of proteins from their primary sequences requires highly automated computational methods linking experimental data. These methods

[‡] Corresponding author

[*] Project (No. 20542006) supported by the National Natural Science Foundation of China

must be able to discriminate the distinct protein function features encapsulated in the protein's structure or in its primary sequences. To this end, the machine learning methods (MLMs) seem to be best suited for the task. Compared to the similarity-based methods such as BLAST or FASTA (Altschul *et al*., 1990) and phylogeny-based method such as ClustalW, MLMs are widely applicable, and now frequently used in biological analysis with relatively good accuracy. MLMs also have a certain degree of flexibility regarding data inputs, allowing them to expand progressively to meet the requirements of rapidly accumulating mountain of data generated from genomics research.

The most often used methods of MLMs are support vector machine (SVM), neural network (NN), hidden markov model (HMM), decision tree (DT) and so on. Among these, SVM is particularly attractive due to its ability to handle noise, large or small datasets, large input spaces (Zavaljevski *et al*., 2002), and its greater accuracy compared to simple BLAST or HMM methods (Karchin *et al*., 2002; Bhasin and Raghava, 2004a; 2004b; Cai *et al*., 2003; 2004).

Currently, there is no reliable systematic way for recognizing and classifying EHs. Barth *et al*.(2004a) reported a method which manually classifies the known EHs into three clusters according to their structural features, that is, the lengths of the loops connecting the core and the cap-loop that is inserted into the cap domain (Table 1). While Barth's method can be used to differentiate certain structural features related to EH's functional features such as substrate specificity, it is hardly useful for newly sequenced proteins due to the large range of sequence similarity between EH sequences.

Strategically, we decided to develop an SVM-based, two-layer, fully automated computational method capable of recognizing EHs first, and then classifying them into their subfamilies based on their protein sequences. We hereby report this work and a user-friendly program EHPred (http://www.wigs.zju.edu/~EHPred) developed on the basis of this study (see RESULTS and DISCUSSION) to assist readers to distinguish EHs and to annotate their subfamilies.

MATERIALS AND METHODS

Datasets

1. Dataset for EH recognition

The sequence data on positive examples of EHs used were obtained from the EH/HD database (Barth *et al*., 2004a) containing 397 protein sequences assigned to three subfamilies according to their structural features such as the lengths of their NC-loops and cap-loops. All proteins denoted as "fragment" or whose annotation was listed as "hypothetical", "similar" or "putative" were removed from the dataset, so that the number of sequences in the dataset was reduced to 231. A non-redundant treatment was applied to eliminate the sequences which share a high degree of similarity (>90%) with others in order to avoid overtraining. The treatment was carried out using the program BLASTCLUST (http://www.ncbi.nlm.nih.gov/BLAST/), which used the BLAST algorithm to systematically cluster protein sequences on the basis of pair-wise matches. The default values were used for all BLAST parameters: matrix BLOSUM62, gap opening cost of 11, gap extension cost of 1, E-value threshold of 1e-6. Under these conditions, the process led to further removal of 92 redundant sequences and reduced the number of sequence sets to 139. These sequences were used as positive examples for EH recognition.

The sequences data on negative examples were obtained from the BRENDA database (Schomburg *et al*., 2004). EH related sequences were removed from the original dataset. A non-redundant treatment was applied (same as for positive datasets) such that no

Table 1 The number of sequences belonging to each EHs subfamily

EH subfamilies	Seq^a	Seq^b	Length of two loops		Comments
			NC-loop	Cap-loop	
Cluster I	36 (72)	16 (43)	16~40	31~59	Cytosolic mammalian and plant EHs and bacterial EHs related to EHs from higher organisms
Cluster II	191 (268)	98 (153)	18~25	5~12	Bacterial EHs
Cluster III	41 (57)	25 (35)	21~57	8~19	Microsomal EHs
Total	268 (397)	139 (231)	–	–	–

Seq^a: Number of original sequences; Seq^b: Number of curated sequences (non-fragment and non-hypothetical sequences). The numbers in parentheses are the corresponding number of sequences before non-redundant treatment sequence had similarity higher than 25% to any others. Thus, 278 non-EH enzyme sequences (two fold of positive examples) were optimized as negative examples.

2. Dataset for classification of subfamilies

The above mentioned 139 sequences of EHs were then grouped into three different clusters (subfamilies) as shown in Table 1. They were used for construction of SVMs for subfamily classification (see below).

Support vector machine

The implementation of SVMs was realized using the software package Gist (Pavlidis *et al*., 2004) which allows us to define a series of parameters as well as the choice of inbuilt kernel functions such as linear, polynomial (of given degree) and radial basis functions (RBF). In the course of this study, all parameters of kernels were kept constant except for the regulatory parameters C and γ which control the trade-off between misclassification error and margin.

Feature vectors extraction

Like other MLMs, SVMs require the data inputs of feature vectors to be of a fixed length, as opposed to the generally variable lengths of protein sequences (Shepherd *et al*., 2003). The extraction of the vector parameters is therefore critical to the performance and accuracy of this method.

In the course of this work, we assessed the use of amino acid composition (AAC), dipeptide composition (DPC) (Reczko and Bohr, 1994), and pseudo-amino acid composition (PAAC), which combines AAC and hydrophobicity and hydrophilicity features along the protein chain (Chou, 2005; Kyte and Doolittle, 1982; Hopp and Woods, 1981), as the feature vectors to encapsulate the global protein information, transforming protein sequences of variable lengths into feature vectors of the required format.

Recognition of an EH

Initially, an SVM module was developed for identifying EHs from diverse protein sequence data. The positive examples were obtained as above described (Table 1). The 278 enzyme sequences from the BRENDA database were used as negative examples (see subsection "Datasets"). The SVM module was trained with AAC, DPC and PAAC as feature vectors respectively. Each feature vector was assessed by linear, polynomial and Radial basis functions (RBFs).

EH subfamily classification

Classifying an unknown EH into a particular subfamily of EH is in fact a multi-class problem. To solve it, three groups of SVMs were constructed in consistence with the three subfamilies of EHs according to Barth's classification. The binary SVMs were designed using a "one-versus-rest" approach (Hua and Sun, 2001), that is, one SVM for each corresponding class (or subset) in the dataset. The SVMs differ from each other by the dataset they are trained with. The *i*th SVM is trained by the data from the *i*th class (subset) as positive and data from all other classes as negative. Each group is comprised of SVMs built with AAC, DPC and PAAC as feature vectors respectively and evaluated by using diverse kernel functions for their performance. These SVMs were used to predict the class of unknown proteins. Inputting an unknown protein sequence into the SVMs yielded three output scores. The highest score from the corresponding SVM indicated the protein's class. The difference between the highest and the second score (referred to as subsection "RI") determines the degree of certainty of the assignment.

Performance evaluation of EHPred

The performance of SVM in distinguishing EHs from non-EHs was evaluated using 5-fold cross-validation. To this end, the datasets were partitioned randomly into five equally sized subsets. The SVMs were then trained with four subsets and tested against the fifth one. The process is reiterated five times so that each subset is used once as the testing data. The performance of EH subfamily classification was evaluated using 2-fold cross-validation because of the relatively low number of sequences.

The evaluation was carried out by measuring the prediction accuracy (PA), the overall prediction accuracy (OPA) (Baldi *et al*., 2000) and Matthews (1975)'s correlation coefficient (MCC). Of these measurements, MCC is considered as a more relevant evaluation as it takes into accounts both over- and under-prediction. In addition, RI (reliability index) is also measured (see below).

Reliability index (RI)

RI is an assessment used to indicate the degree of confidence in the prediction to the user when using machine-learning techniques (Reinhardt and Hubbard, 1998; Emanuelsson *et al*., 2000). RI is determined according to the difference (Δ) between the highest and the second highest value the SVMs gave in multi-class classification. The higher the RI is, the greater the probability that the prediction is accurate. In this study, RI is determined as follows:

$$RI=\begin{cases}1 & \text{if } 0\le\Delta<0.5\\ INT(\Delta)\times2+1 & \text{if } 0.5\le\Delta<2.5\\ 5 & \text{if } \Delta\ge2.5\end{cases}\qquad(1)$$

RESULTS AND DISCUSSION

Results of evaluation of EH recognition are summarized in Table 2. An EH sequence can be distinguished from other protein sequences with an accuracy of 94.2% and an MCC of 0.89. The best results were obtained using DPC with an RBF kernel (γ=4, C=1) in our analysis. This type of SVM was also compared with BLAST and NN (neural network) method (The structure of NN used in our analysis is illustrated in Fig.S1[1], see http://www.wigs.zju.edu.cn/~EHPred/pdf/EHPred_Supplementary_Material.pdf). It gave clearly better performance compared to BLAST (86%) and NN (80%) methods (Table S3, see http://www.wigs.zju.edu.cn/~EHPred/pdf/EHPred_Supplementary_Ma-terial.pdf). This is not surprising because the BLAST method can hardly provide reliable results of fair accuracy when the sequence identities between the proteins were low. NN method is also less accurate since it usually requires large datasets. It is known that EHs of different origins with a highly conserved structure may differ from each other in a large range of sequences identity (as low as 17% in some cases) (Barth *et al.*, 2004b). The results suggest the SVM based methods are the most effective in such situations.

The best results of evaluation for EH subfamily classification are obtained using a RBF kernel with γ=16 and C=1. PAAC performed the best with an overall accuracy of 90.7% as compared to AAC (80.0%) and DPC (84.9%). The detailed dataset partition for the 2-fold cross validation and the results of the predictions along with their accuracy under various parameters are provided in the supplementary materials section (http://www.wigs.zju.edu.cn/~EHPred/pdf/EHPred_Supplementary_Material.pdf). The relationship between the expected prediction accuracy and RI of each method (DPC and PAAC) are presented in Fig.1. The fraction of sequences and their corresponding prediction accuracy at a given RI are also given in Table 3. In terms of distribution of the number of sequences in our dataset along the value of RI, about 44.6% of the sequences had RI of 5, which showed that 71.9% of sequences can be predicted with $RI\ge3$ and that these sequences were 90.9% correctly predicted based on PAAC. The prediction is considered reliable in our program when $RI\ge3$.

As summarized in Table 2, PAAC gave better results as compared to AAC and DPC in EH subfamily classification, especially of Class I and Class II subfamilies which are relatively small in size. PAAC contains 20+2λ discrete numbers for different training dataset. λ has an optimal value for each different dataset. In our study, the optimal value for λ was 30. The results revealed that the subfamilies of EHs are different in distributions of hydrophobicity and hydrophilicity within amino acid residues along a protein chain. Further investigation revealed that the cap-loop of EH from *Aspergillus niger* (subfamily III) contains 25% hydrophobic residues while EH from *Mus musculus* (subfamily I) is comprised of 63% hydrophobic residues within the same region (Nardini *et al.*, 1999; Zou *et al.*, 2000; Argiriadi *et al.*, 1999). This characteristic appears to be an important factor in determining the protein's structure and function.

In order to further correlate the difference between subfamilies of EHs to PAAC, the informational entropy (Shannon entropy) of the amino acid sequences of EHs was calculated during the multiple sequence alignment. Hydrophobic amino acids such as Gly and Pro are frequently recognized as conservative residues ($H_j\approx0$) (Table 4) (Varfolomeev *et al.*, 2002). The same profiles were also shown in the catalytic motif logos (Fig.S2, see http://www.wigs.zju.edu.cn/~EHPred/pdf/EHPred_Supplemetary_Material.pdf) where Gly and Pro were residues with high frequency. It is known that Pro can act as a structural disruptor for (α) helices and a turning point in β sheets, which causes particular bending of the protein chain and inhibits the formation of close-packed, ordered secondary structures that are vital in providing a solid

[1] Note: Tables and figures in supplementary file are numbered with an S before the numbers, i.e. Table S1, Fig.S1, differing these in main text, Table 1, Fig.1

Table 2 The performance of our method in differentiating EHs from non-EHs and identifying the three subfamilies of EHs

Feature	EHs recognition (RBF with γ=4, C=1)		EHs classification (RBF with γ=16, C=1)						
			Cluster I		Cluster II		Cluster III		Overall
	ACC (%)	*MCC*	*ACC* (%)	*MCC*	*ACC* (%)	*MCC*	*ACC* (%)	*MCC*	*ACC* (%)
AAC	89.21	0.77	62.50	0.43	81.63	0.61	80.00	0.56	80.04
DPC	**94.24**	**0.89**	62.50	0.77	91.84	0.73	72.00	0.82	84.89
PAAC (λ=30)	82.45	0.74	**81.25**	**0.83**	**92.86**	**0.78**	**88.00**	**0.84**	**90.65**

ACC: Accuracy; MCC: Matthews's correlation coefficient; AAC: Amino acid composition; DPC: Dipeptide composition; PAAC: Pseudo-amino acid composition (λ=30). Radial basis function (RBF) kernel has been used to construct to classifiers. The parameter γ was adaptively adjusted to 4 in EHs recognition and 16 in EHs classification respectively. Regulatory parameter C was optimized to 1. EHs recognition results were obtained through 5-fold cross-validation, EHs classification evaluated by 2-fold cross-validation. Boldface indicates the best performance

Table 3 The fractions of sequences and prediction accuracy at a given RI based on different types of compositions

RI	*AAC*		*DPC*		*PAAC*	
	Frequency (%)	*ACC* (%)	Frequency (%)	*ACC* (%)	Frequency (%)	*ACC* (%)
1	10.07	7.14	15.11	28.57	15.10	28.60
2	7.91	36.36	21.58	90.00	12.90	77.88
3	7.91	45.46	25.90	91.67	15.80	90.90
4	9.35	53.85	21.58	96.67	11.50	100
5	64.75	74.44	15.83	100	44.60	100

AAC: Amino acid composition; DPC: Dipeptide composition; PAAC: Pseudo-amino acid composition

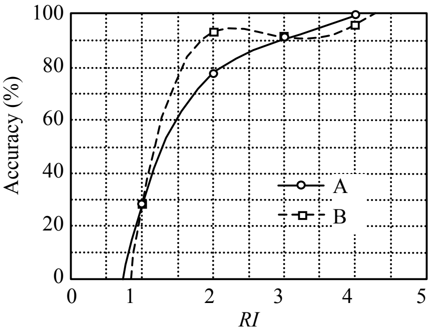

Fig.1 The relationships between the expected prediction accuracy and RI of each method
A: Expected prediction accuracy of a PAAC-based classifier with RI equal to a given value; B: Expected prediction accuracy of a DPC-based classifier with RI equal to a given value

connection with the substrates. Gly can improve the conformational flexibility of the active site because of its small side chain and energetically favorable rotation along the C-N and C-C bonds of the polypeptide chain. They contribute to the formation of the active site architecture. The preferred hydrophobic amino acids such as Gly, Ala, Ile, Leu, Phe, Pro, Trp and Val can be seen beside the catalytic residues in the catalytic motif logos.

The special amphiphilic features of the catalytic motif and loop regions reflect the substrate's specificity of different subfamilies of EHs. This allows the PAAC feature vector to be especially effective.

Table 4 The informational entropy of the EHs

Amino acid	Frequency (%)
G	24.59
P	16.39
H	9.84
W	8.20
D	6.56
F	6.56
A	3.28
L	3.28
N	3.28
Q	3.28
S	3.28
Others	11.48

It is expected that the accuracy in prediction of EHs subfamilies and the feature they share can be further improved as additional protein sequences and experimental data become available for use in training. In the future, further improvements rely on increased quantity and quality of datasets with more experimental parameters, and comprehensive information on proteins.

CONCLUSION

In conclusion, this paper describes an effective SVM-based method for recognizing and classifying EHs with high accuracy. The program, EHPred, devised on the basis of this study, can assist in the annotation of the sequence data and assignment of subfamilies of EHs produced in ongoing sequencing projects. This method may find wide applications in reducing the gap between the rapidly increasing amount of genome sequence data collected and their annotation rate. Ultimately, it may facilitate drug discovery related to EHs and their many associated diseases and drug metabolisms.

A supplementary file provides more details of the methodology and results that are not covered in the main text and can be downloaded from our website: http://www.wigs.zju.edu.cn/~EHPred/pdf/EHPred_Supplemetary_Material.pdf.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**:403-410. [doi:10.1006/jmbi.1990.9999]

Argiriadi, M.A., Morisseau, D., Hammock, B.D., Christianson, D.W., 1999. Detoxification of environmental mutagens and carcinogens: structure, mechanism, and evolution of liver epoxide hydrolase. *Proceedings of the National Academy of Sciences USA*, **96**:10637-10642. [doi:10.1073/pnas.96.19.10637]

Armstrong, R.N., 1987. Enzyme-catalyzed detoxication reactions: mechanisms and stereochemistry. *CRC Critical Reviews in Biochemistry*, **22**:39-88.

Baldi, P., Brunak, S., Chauvin, Y., Anderson, C.A.F., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**:412-419. [doi:10.1093/bioinformatics/16.5.412]

Barth, S., Fischer, M., Schmid, R.D., Pleiss, J., 2004a. The database of epoxide hydrolases and haloalkane dehalogenases: one structure, many functions. *Bioinformatics*, **20**:2845-2847. EH/HD database can be available at http://www.led.uni-stuttgart.de. [doi:10.1093/bioinformatics/bth284]

Barth, S., Fischer, M., Schmid, R.D., Pleiss, J., 2004b. Sequence and structure of epoxide hydrolases: a systematic analysis. *PROTEINS: Structure, Function, and Bioinformatics*, **55**:846-855. [doi:10.1002/prot.20013]

Bhasin, M., Raghava, G.P.S., 2004a. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic. Acids Research*, **32**:W383-W389. [doi:10.1093/nar/gkh001]

Bhasin, M., Raghava, G.P.S., 2004b. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic. Acids Research*, **32**:W414-W419.

Cai, C.Z., Wang, W.L., Sun, L.Z., Chen, Y.Z., 2003. Protein function classification via support vector machine approach. *Mathematical Biosciences*, **185**:111-122. [doi:10.1016/S0025-5564(03)00096-8]

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, Y.Z., 2004. Enzyme family classification by support vector machines. *PROTEINS: Structure, Function, and Bioinformatics*, **55**:66-76. [doi:10.1002/prot.20045]

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**:10-19. [doi:10.1093/bioinformatics/bth466]

Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300**:1005-1016. [doi:10.1006/jmbi.2000.3903]

Fretland, A.J., Omiecinski, C.J., 2000. Epoxide hydrolases: biochemistry and molecular biology. *Chemico-Biological Interactions*, **129**:41-59. [doi:10.1016/S0009-2797(00)00197-6]

Hopp, T.P., Woods, K.R., 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences USA*, **78**:3824-3828.

Hua, S.J., Sun, Z.R., 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**:721-728. [doi:10.1093/bioinformatics/17.8.721]

Karchin, R., Karplus, K., Haussler, D., 2002. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**:147-159. [doi:10.1093/bioinformatics/18.1.147]

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**:105-132. [doi:10.1016/0022-2836(82)90515-0]

Matthews, B.W., 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochimica. Biophysica. Acta*, **405**:442-451.

Nardini, M., Ridder, I.S., Rozeboom, H.J., Kalk, K.H., Rink, R., Janssen, D.B., Dijkstra, B.W., 1999. The X-ray structure of epoxide hydorlase from *Agrobacterium radiobacter* AD1. *Journal of Biological Chemistry*, **274**:14579-14586. [doi:10.1074/jbc.274.21.14579]

Pavlidis, P., Wapinski, I., Noble, W.S., 2004. Support vector machine classification on the web. *Bioinformatics*, **20**:586-587. [doi:10.1093/bioinformatics/btg461]

Reczko, M., Bohr, H., 1994. The DEF data base of sequence based protein fold class predictions. *Nucleic. Acids Research*, **22**:3616-3619.

Reinhardt, A., Hubbard, T., 1998. Using neural networks for prediction of the subcellular location of proteins. *Nucleic. Acids Research*, **26**:2230-2236. [doi:10.1093/nar/26.9.2230]

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D., 2004. BRENDA, the enzyme database: updates and major new developments. *Nucleic. Acids Research*, **32**:D431-D433. The BRENDA database can be available at http://www.brenda.uni-koeln.de/. [doi:10.1093/nar/gkh081]

Shepherd, A.J., Gorse, D., Thornton, J.M., 2003. A novel approach to the recognition of protein architecture from sequence using Fourier analysis and neural networks. *Proteins*, **50**:290-302. [doi:10.1002/prot.10290]

Varfolomeev, S.D., Uporov, I., Fedorov, E.V., 2002. Bioinformatics and molecular modeling in chemical enzymology active sites of hydrolases. *Biochemistry (Moscow)*, **67**:1328-1340. [doi:10.1023/A:1020907122341]

Zavaljevski, N., Stevens, F.J., Reifman, J., 2002. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**:689-696. [doi:10.1093/bioinformatics/18.5.689]

Zou, J., Hallberg, B.M., Bergfors, T., Oesch, F., Arand, M., Mowbray, S.L., Jones, T.A., 2000. Structure of *Aspergillus niger* epoxide hydrolase at 1.8 A resolution: implications fro the structure and function of the mammalian microsomal class of epoxide hydrolases. *Structure*, **8**:111-122. [doi:10.1016/S0969-2126(00)00087-3]