SUPPORT VECTOR MACHINES FOR PREDICTING APOPTOSIS PROTEINS TYPES

Jing Huang[1] and Feng Shi[2,3,*]

[1]School of Computer, Wuhan University, Wuhan, Hubei Province, PR China.
E-mail: caroline_hj@sohu.com
[2]School of Science, Huazhong Agricultural University, Wuhan, Hubei Province, PR China.
[3]School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei Province, PR China. E-mail: shifeng@mail.hzau.edu.cn
*Corresponding author

Received 15 January 2004; revised 17 May 2004; accepted 7 October 2004

ABSTRACT

Apoptosis proteins have a central role in the development and homeostasis of an organism. These proteins are very important for understanding the mechanism of programmed cell death, and their function is related to their types. According to the classification scheme by Zhou and Doctor (2003), the apoptosis proteins are categorized into the following four types: (1) cytoplasmic protein; (2) plasma membrane-bound protein; (3) mitochondrial inner and outer proteins; (4) other proteins. A powerful learning machine, the Support Vector Machine, is applied for predicting the type of a given apoptosis protein by incorporating the sqrt-amino acid composition effect. High success rates were obtained by the re-substitute test ($98/98 = 100\ \%$) and the jackknife test ($89/98 = 90.8\%$).

Key Words: support vector machine, subcellular location, sqrt-amino acid composition

1. INTRODUCTION

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death (e.g. Jacobson *et al*., 1997; Chou *et al*., 1999; Chou and Maggiora, 1998; Chou *et al*., 1997; Chou *et al*., 2000; Zhou *et al*., 1999). This process entails the autolytic degradation of cellular components, and is characterized by blebbing of cell membranes, shrinkage of cell volumes, and condensation of nuclei (Kerr *et al*., 1972), and is currently an area of intense investigation. Cell death and renewal are responsible for maintaining the proper turnover of cells, which ensures a constant controlled flux of fresh cells. Programmed cell death and cell proliferation are tightly coupled. When apoptosis malfunctions, a variety of formidable diseases can ensue: blocking apoptosis is associated with cancer (Evan and Littlewood, 1998) and autoimmune disease, whereas unwanted apoptosis can possibly lead to ischemic damage (Reed and Paternostro, 1999) or neurodegenerative disease (Schulz *et al*., 1999). Apoptosis is considered to have a key role in these several devastating diseases and, in principle, provides many targets for therapeutic intervention (Barinaga, 1998; Chou *et al*., 1997; Chou *et al*., 2000).

To understand the apoptosis mechanism and functions of various apoptosis proteins, it would be helpful to obtain information about their subcellular location. This is because the subcellular location of apoptosis proteins is closely related to their function (Chou, 2000a; Chou and Cai, 2003a; Chou and Elrod, 1999a, 1999b; Reed and Paternostro, 1999; Suzuki *et al*., 2000). It has been known that there are 732 archetypical proteins with "apoptosis" domains (Zhou and Doctor, 2003), and only 98 of these proteins are known to be the apoptosis protein (for more details, one can visit: http://www.apoptosis-db.org). Scientists usually deal with a number of protein sequences already known belonging to apoptosis proteins. However, it is both time-consuming and costly to determine which specific subcellular location a given apoptosis protein belongs to. Confronted with such a situation, can we develop a fast and effective way to predict the subcellular location for a given apoptosis protein based on its amino acid sequence? Recently, Guo-ping Zhou (Zhou and Doctor, 2003) attempted to identify the subcellular location of apoptosis proteins according to their sequences by means of the covariant discriminant function, which was established based on the Mahalanobis distance and Chou's invariance theorem (Chou, 1995; Chou, 2000a; Chou, 2001; Chou and Zhang, 1995; Pan *et al*., 2003; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). The results were quite promising, indicating that the subcellular location of apoptosis proteins are predictable to a considerably accurate extent if a good vector representation of protein can be established. It is expected that, with a continuous improvement of vector representation methods by incorporating amino acid properties, some predicting method might eventually become a useful tool in this area because the function of an apoptosis protein is closely related to its subcellular location. The present study was initiated in an attempt to address this problem.

Chou and Elrod performed extensive research in predicting subcellular location mainly based on amino acid composition (Chou and Elrod, 1999a, 1999b). Subsequently, in order to take into account the sequence-order effects and improve the prediction quality, Chou has further incorporated the quasi-sequence order effect (Chou, 2001) and introduced the concept of "pseudo-amino-acid composition" (Chou, 2001; Pan *et al*., 2003). For example, Chou (1999) classified membrane proteins into five different types and proposed a covariant discriminant algorithm to predict the types of membrane proteins. Recently, Cai *et al*. (2001) applied neural networks to this problem. To improve the prediction quality, Chou and Cai (2002) proposed a new method in which the covariant discriminate algorithm was augmented to incorporate the quasi-sequence-order effect. This method uses the amino acid composition and the sequence-order-coupling numbers (reflecting the sequence order effect) in order to improve the prediction quality. Feng (2001) proposed a new representation of unified attribute vector explaining that each protein can be represented by a vector, which is a 20-D vector in Hilbert space with unified length. Hence, all proteins have their representative points on the surface of the 20-D globe. The representative points of the proteins in the same family or with the higher sequence identity are closer on the surface. The overall predictive accuracy could be improved from 3 to 5% for different databases (Feng, 2001) with this simply modification of the usage of the amino acid composition. Recently, a series of new powerful approaches have been developed by Chou and his co-workers (Cai and Chou, 2003; Cai *et al*., 2003; Chou and Cai, 2002; Chou and Cai, 2003a, 2003b). Encouraged by the great successes of the previous investigators in the area, here we would like to use a different strategy, involving support vector machines, to approach predication of

apoptosis protein types in hope that our approach can play a complementary role to the existing methods.

2. SUPPORT VECTOR MACHINE

A support vector machine (SVM) is one type of learning machines based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated briefly as follows.

Suppose we are given a set of samples, i.e. a series of input vectors $X_i \in R^d$ ($i = 1, 2, \ldots, N$), with corresponding labels $y_i \in \{+1, -1\}$ ($i = 1, 2, \ldots, N$) , where –1 and +1 are used to stand respectively for the two classes. The goal here is to construct one binary classifier or derive one decision function from the available samples, which has small probability of misclassifying a future sample. Only the most useful linear non-separable case (for most real life problems) are considered here.

SVM performs a nonlinear mapping of the input vector x from the input space R^d into a higher dimensional Hilbert space, where the mapping is determined by the kernel function. It finds the OSH (Optimal Separating Hyperplane, see Cortes and Vapnik, 1995) in the space H corresponding to a non-linear boundary in the input space. Two typical kernel functions are listed below:

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + 1)^d, \tag{1}$$

$$K(\vec{x}_i, \vec{x}_j) = \exp(-\lambda ||\vec{x}_i - \vec{x}_j||^2). \tag{2}$$

The first one is called the polynomial kernel function of degree d which will eventually revert to the linear function when $1 = d$, and the latter one is called the RBF (radial basis function) kernel with one parameter λ. Finally, for the selected kernel function, the learning task amounts to solving the following convex quadratic programming (QP) problem,

$$\text{Max} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(\vec{x}_i \cdot \vec{x}_j)$$

subject to:

$$0 \leq \alpha_i \leq C$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

where the form of the decision function is

$$f(\vec{x}) = \text{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{x}_i \cdot \vec{x}_j) + b \right).$$

For a given data set, only the kernel function and the regularity parameter C must be selected.

A complete description to the theory of SVMs for pattern recognition is found in Vapnik (1998). SVMs have been used in a range of bioinformatics problems including protein fold recognition (Ding and Dubchak, 2001); protein–protein interactions prediction (Cai *et al*., 2000a, 2000b); prediction of protein subcellular location

(Cai *et al*., 2000, 2002; Hua and Sun, 2001a), protein secondary structure prediction (Hua *et al*., 2001b), *T*-cell epitopes prediction (Zhao *et al*., 2003), classification of protein quaternary structure (Zhang *et al*., 2003).

In this paper, we apply Vapnik's Support Vector Machine for predicting the types of apoptosis proteins. We have used the OSU_{SVM}, a Matlab SVM toolbox (can be download freely from: http://www.ece.osu.edu/∼maj/osu_svm), which is an implementation of SVM for the problem of pattern recognition.

3. TRAINING AND PREDICTION

According to their subcellular location (Zhou and Doctor, 2003), apoptosis proteins are classified into the following four types: (1) type I: Cytoplasmic protein; (2) type II: Plasma membrane-bound protein; (3) type III: Mitochondrial inner and outer proteins; (4) type IV:Other proteins.

In this research, every protein is represented as a point or a vector in a 20-D space. Every component of its vector was supposed to be the square root of occurrence frequencies of the 20 amino acids in the protein concerned. Therefore, we used $\sqrt{f_i}$ to replace f_i as the input of the SVM. As an example, the 20-D vector of occurrence frequencies of the 20 amino acids of the protein NP_001151 can be computed as: (0.0218, 0.0327, 0.0218, 0.0360, 0.0235, 0.0402, 0.0461, 0.0427, 0.0519, 0.0394, 0.0695, 0.0243, 0.0494, 0.0737, 0.0637, 0.0695, 0.0796, 0.0494, 0.0528, 0.1122), and the represented vector was supposed to be the square root of occurrence frequencies:(0.1476, 0.1807, 0.1476, 0.1898, 0.1531, 0.2005, 0.2146, 0.2067, 0.2279, 0.1984, 0.2637, 0.1558, 0.2223, 0.2715, 0.2523, 0.2637, 0.2821, 0.2223, 0.2297, 0.3350). In this article, we call this represented vector as the *sqrt-amino acid composition*. The *sqrt-amino acid compositions* have the property that they all are unit vectors of the 20-D Hilbert space, hence, all proteins have their representative points on the surface of the 20-D globe. We expect that the representative points of the proteins in the same family are closer on the surface, while the core of the SVM is mapping the lower dimensional vectors into a higher dimensional Hilbert, so that the these vectors can be classified OSH in the Hilbert. That is the reason we choose the *sqrt-amino acid composition*. At the same time, we also standard the sqrt-amino acid composition to unit length (which is no longer an unit vector), and get almost the same high accuracy in total.

The computations were carried out on a PC. We have done many experiments by using difference kernel function (including linear, polynomial and Gaussian RBF kernel functions) as well as difference parameters, and find that by using Gaussian RBF kernel function, the performance of the SVM is best with a proper parameter. For the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension. After being trained, the hyper-plane output by the SVM was obtained. The SVM method applies to two-class problems. In this paper, for the four-class problems, we have used a simple and effective method: "one-against-others" method (Ding and Dubchak, 2001) to transfer it into two-class problems. We first test the self-consistency and leave-one-out cross-validation (jackknife test) of the method, followed by testing the method by prediction of an independent dataset. As a result, the rates of self-consistency, cross-validation of prediction were quite high.

In addition to the prediction algorithm, we also need to construct a training data set to complete the establishment of a statistical prediction method. To realize this, based

Table 1. List of the accession numbers for the 98 apoptosis proteins classified into four categories according to their subcellular location[a].

Type I: 43 cytoplasmic proteins							
XP_013050	P55212	P42574	P39429	P55867	P55865	Q02357	NP_033941
NP_033940	NP_033939	NP_031637	NP_031570	NP_031563	NP_031490	NP_033447	P29452
NP_036246	NP_001218	NP_004041	O54786	Q60989	Q62210	NP_065209	NP_001151
NP_071610	NP_071567	NP_066961	NP_037054	NP_036894	NP_005649	NP_004392	NP_004315
NP_001187	NP_001159	NP_001157	NP_001156	P22366	P55866	Q60431	P55214
P55269	P29466	O70201					
Type II: 30 plasma membrane-bound proteins							
NP_037223	P28825	NP_037275	NP_032013	NP_032612	P50555	P25118	P18519
O19131	Q63199	O77736	P51867	NP_036742	NP_037315	NP_005916	NP_005579
NP_000034	NP_001056	NP_003781	NP_002498	O02703	Q13014	NP_031553	NP_031549
Q63690	Q07820	NP_001179	Q91828	Q91827	Q07812		
Type III: 13 mitochondrial inner and outer proteins							
XP_008738	O77737	Q00709	NP_033873	P10417	P53563	Q07816	P49950
Q07817	O95831	Q9OX1	Q9JM53	Q9VQ79			
Type IV: 12 other proteins[b]							
Q63369	Q90660	Q00653	Q04861	P19838	NP_032715	P98150	Q15121
Q62048	NP_033872	NP_004040	NP_005736				

[a]Derived from SWISS-PROT data bank.
[b]Of the 12 other apoptosis proteins, five are located in nucleus, two in endoplasmic reticulum, one in microtubule, and one in lysosome (Zhou and Doctor, 2003).

on the SWISS-PROT data bank, 98 apoptosis proteins (the dates were taken from Zhou and Doctor (2003)) were classified into the following four subcellular locations: (1) cytoplasmic, (2) plasma membrane-bound, (3) mitochondrial, and (4) other (Table 1).

4. RESULTS AND DISCUSSION

By means of the SVM algorithm described in the last section, a statistical prediction was performed for the 98 apoptosis proteins listed in Table 1. The prediction was conducted by two different approaches, the re-substitution test and the jackknife test. The results are given in Table 2.

Re-substitution test

The so-called re-substitution test is an examination for the self-consistency of a prediction method (Chou and Maggiora, 1998; Chou and Zhang, 1994; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). When the re-substitution test was performed for the current study, the type of each apoptosis protein in a data set was in turn identified using the rule parameters derived from the same data set, the so-called training data set. As shown in Table 2, the overall success rate thus obtained for the 98 apoptosis proteins in Table 1 was 100%, indicating an excellent self-consistency.

Table 2. Tested Results for the 98 apoptosis proteins in Table 1 by both.

	Success rate				
Test method	Type I	Type II	Type III	Type IV	Overall
	Re-substitution test and jackknife test				
Re-substitute					
Covariant[a]	43/43 = 100%	30/30 = 100%	9/13 = 60.2%	7/12 = 58.3%	89/98 = 90.8%
SVM I[b]	42/43 = 97.7%	30/30 = 100%	13/13 = 100%	12/12 = 100%	97/98 = 99.0%
SVM II[c]	43/43 = 100%	30/30 = 100%	13/13 = 100%	12/12 = 100%	98/98 = 100%
SVM III[d]	42/43 = 97.70%	30/30 = 100%	13/13 = 100%	12/12 = 100%	97/98 = 99.0%
Jack-knife					
Covariant[a]	42/43 = 97.7%	22/30 = 73.3%	4/13 = 30.8%	3/12 = 25.0%	71/98 = 72.5%
SVM I[b]	36/43 = 83.5%	26/30 = 86.7%	12/13 = 92.5%	9/12 = 75.0%	83/98 = 84.8%
SVM II[c]	37/43 = 86.0%	27/30 = 90.0%	13/13 = 100%	12/12 = 100%	89/98 = 90.8%
SVM III[d]	39/43 = 91.4%	28/30 = 93.3%	12/13 = 92.5%	9/12 = 75.0%	88/98 = 89.8%

[a]Comes from Guo-Ping Zhou and Doctor (2003), by using covariant discriminant function.
[b]Using SVM by occurrence frequencies of the 20 amino acids.
[c]Using SVM by sqrt-amino acid composition.
[d]Using SVM by standarding the aqrt-aminoacid composition to unit length.
All use Gauss RBF kernel function, while the value C were set to 100, 20, 200 according to a, b and c respectively.

However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to derive the rule parameters and to test themselves. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of a prediction method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of a prediction method in practical application. This is important especially for checking the validity of a training data set—whether it contains sufficient information to reflect all the important features concerned so as to field a high success rate in application.

Jackknife test

As is well known, the independent data set test, sub-sampling test, and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one (see Chou and Zhang, 1995) for a comprehensive discussion about this). During jackknifing, each protein in the data set is in turn singled out as a tested protein and all the rule parameters are calculated based on the remaining proteins. In other words, the

subcellular location of each apoptosis protein is identified by the rule parameters derived using all the other apoptosis proteins except the one that is being identified. During the process of jackknifing, both the training data set and testing data set are actually open, and a protein will in turn move from one to the other. As expected, the success prediction rates by jackknife test were decreased compared with those by the re-substitution test. Such a decrement is particularly more remarkable for small subsets. This is because the cluster-tolerant capacity (Chou, 1999) for small subsets is usually low. And hence the information loss resulting from jackknifing will have a greater impact on the small subsets than the large ones. Nevertheless, as shown in Table 2, the overall jackknife rate for the data set of the 98 apoptosis proteins could still reach 90.8%. It is expected that the success rate for identifying the subcellular location of apoptosis proteins can be further enhanced by improving the training data of small subsets by adding into them more new proteins that have been found belonging to the subcellular location defined by these subsets.

5. CONCLUSIONS

The above results, together with those obtained by the covariant discriminant prediction algorithm (Chou, 2000a, 2000b; Chou and Cai, 2002), have indicated that the types of apoptosis proteins are predictable with a considerable accuracy. It is anticipated that the covariant discriminant algorithm (Chou and Cai, 2003a), and the SVM, if effectively complemented with each other, will become a powerful tool for predicting the types of apoptosis proteins. The current study has further demonstrated that the datasets originally constructed by Zhou and Doctor (2003) will be very useful for the area of apoptosis study. It is expected that the prediction quality can be further improved if the current SVM can be properly combined with pseudo–amino acid composition (Chou, 2001; Pan *et al*., 2003) and function domain composition (Cai *et al*., 2003; Chou and Cai, 2002) and with other amino acid properties.

ACKNOWLEDGMENTS

The authors thank Dr. G.P. Zhou for providing the amino acid sequences of the apoptosis proteins and some helpful discussions, thank the referees and the editor for many useful suggestions. The work was partly supported by the National Natural Science Funds of China (NO: 30170214), and Huazhong Agricultural University.

REFERENCES

Barinaga, M. (1998). Stroke-damaged neurons may commit cellular suicide. Science 281: 1302–1303.

Cai, Y.D. and K.C. Chou (2003). Nearest neighbour algorithm for predicting protein subcellular location by combing functional domain composition and pseudo-amino acid composition. Biochemical and Biophysical Research Communications 305: 407–411.

Cai, Y.D., G.P. Zhou and K.C. Chou (2003). Support vector machine for predicting membrane protein types by using functional domine composition. Biophysical Journal 84: 3257–3263.

Cai, Y.D., X.J. Liu and K.C. Chou (2001). Artificial neural network model for predicting membrane protein types. Journal of Biomolecular Stereodynamics 18: 607–610.

Cai, Y.D., X.J. Liu, X.B. Xu and K.C. Chou (2000a). Support vector machines for prediction of protein subcellular location. Molecular Cell Biology Research Communications 4: 230–233.

Cai, Y.D., X.J. Liu, X.B. Xu and K.C. Chou (2000b). Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. Journal of Biological Chemistry 84: 343–348.

Chou, J.J., H. Li, G.S. Salvesen, J. Yuan and G. Wagner (1999). Solution structure of BID, an intracellular amplifier of apoptotic signaling. Cell 96: 615–624.

Chou, K.C. (1995). A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function and Genetics 21: 319–344.

Chou, K.C. (1999). A key driving force in determination of protein structural classes. Biochemical and Biophysical Research Communications 264: 216–224.

Chou, K.C. (2000a). Prediction of protein subcellular locations by incorporation quasi-sequence-order effect. Bichemical and Biophysical Research Communication 19: 477–483.

Chou, K.C. (2000b). Review: Prediction of protein structural classes and subcellular location. Current Protein and Peptide Science 1: 171–208.

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

Chou, K.C. and Y.D. Cai (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. Journal of Biology Chemistry 277: 45765–45769.

Chou, K.C. and Y.D. Cai (2003a). A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology. Bichemical and Biophysical Research Communication 311: 743–747.

Chou, K.C. and Y.D. Cai (2003b). Predication and classification of protein subcellular location:sequence-order effect and pseudo amino composition. Journal of Cellular Biochemistry 90: 1250–1260.

Chou, K.C. and D.W. Elrod (1999a). Prediction of membrane protein types and subcellular locations. Proteins 34: 137–153.

Chou, K.C. and D.W. Elrod (1999b). Protein subcellular location prediction. Protein Engineering 12: 107–118.

Chou, K.C., D. Jones and R.L. Heinrikson (1997). Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Letters 419: 49–54.

Chou, K.C. and G.M. Maggiora (1998). Domain structural class prediction. Protein Engineering 11: 523–538.

Chou, K.C., A.G. Tomasselli and R.L. Heinrikson (2000). Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Letters 470: 249–256.

Chou, K.C. and C.T. Zhang (1994). Review: Predicting protein folding types by distance functions that make allowances for amino acid interactions. Journal of Biological Chemistry 269: 22014–22020.

Chou, K.C. and C.T. Zhang (1995). Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30: 275–349.

Cortes, C. and V. Vapnik (1995). Support vector networks. Machine Learning 20: 273–293.

Ding, C.H. and I. Dubchak (2001). Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17: 349–358.

Evan, G. and T. Littlewood (1998). A matter of life and cell death. Science 281: 1317–1322.

Feng, Z.P. (2001). Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499.

Hua, S.J. and Z.R. Sun (2001a). Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728.

Hua, S.J. and Z.R. Sun (2001b). A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. Journal of Molecular Biology 308: 397–407.

Jacobson, M.D., M. Weil and M.C. Raff (1997). Programmed cell death in animal development. Cell 88: 347–354.

Kerr, J.F., A.H. Wyllie and A.R. Currie (1972). Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics. British Journal of Cancer 26: 239–257.

Pan, Y.X., Z.Z. Zhang, Z.M. Guo, G.Y. Feng, Z.D. Huang and L. He (2003). Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach. Journal of Protein Chemistry 22: 395–402.

Reed, J.C. and G. Paternostro (1999). Postmitochondrial regulation of apoptosis during heart failure. Proceedings of the National Academy of Sciences 96: 7614–7616.

Schulz, J.B., M. Weller and M.A. Moskowitz (1999). Caspases as treatment targets in stroke and neurodegenerative diseases. Annals of Neurology 45: 421–429.

Suzuki, M., R.J. Youle and N. Tjandra (2000). Structure of Bax: Coregulation of dimer formation and intracellular location. Cell 103: 645–654.

Vapnik, V. (1998). Statistical Learning Theory. Wiley Interscience, New York.

Zhang, S.W., Q. Pan, H.C. Zhang, Y.L. Zhang and H.Y. Wang (2003). Classification of protein quaternary structure with support vector machine, Bioinformatics 19: 2390–2396.

Zhao, Y.D., C. Pinilla, D. Valmori, R. Martin and R. Simon (2003). Application of support vector machines for *T*-cell epitopes prediction, Bioinformatics 19: 1978–1984.

Zhou, G.P. (1998). An intriguing controversy over protein structural class prediction. Journal of Protein Chemistry 17: 729–738.

Zhou, G.P. and N. Assa-Munt (2001). Some insights into protein structural class prediction. Proteins: Structure, Function, and Genetics 44: 57–59.

Zhou, G.P. and K. Doctor (2003). Subcelluar location of Apoptosis proteins. Proteins: Structure, Function, and Genetic 50: 44–48.

Zhou, P., J. Chou, R.S. Olea, J. Yuan and G. Wagner (1999). Solution structure of Apaf-1 CARD and its interaction with caspase-9 CARD: A structural basis for specific adaptor/caspase interaction. Proceedings of the National Academy of Sciences 96: 11265–11270.