Homology-Free Prediction of Functional Class of Proteins and Peptides by Support Vector Machines

F. Zhu[1,4], L.Y. Han[1], X. Chen[2], H.H. Lin[1], S. Ong[1], B. Xie[1], H.L. Zhang[1] and Y.Z. Chen[1,3,*]

[1]*Bioinformatics and Drug Design Group, Department of Pharmacy and Center for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543;* [2]*Department of Biotechnology, Zhejiang University, P.R. China, 310029;* [3]*Shanghai Center for Bioinformatics Technology, Shanghai, 201203, P. R. China;* [4]*Department of Biological Science, National University of Singapore, 14 Science Drive 4, Singapore 117543*

Abstract: Protein and peptide sequences contain clues for functional prediction. A challenge is to predict sequences that show low or no homology to proteins or peptides of known function. A machine learning method, support vector machines (SVM), has recently been explored for predicting functional class of proteins and peptides from sequence-derived properties irrespective of sequence similarity, which has shown impressive performance for predicting a wide range of protein and peptide classes including certain low- and non- homologous sequences. This method serves as a new and valuable addition to complement the extensively-used alignment-based, clustering-based, and structure-based functional prediction methods. This article evaluates the strategies, current progresses, reported prediction performances, available software tools, and underlying difficulties in using SVM for predicting the functional class of proteins and peptides.

Keywords: Machine learning method, peptide, peptide function, protein family, protein function, protein function prediction, protein sequence, support vector machine.

INTRODUCTION

Amino acid sequence of proteins and peptides contain clues for predicting their functions [1-7]. Sequence similarity [8-10], motifs [11], clustering [12-14], and evolutionary relationships [15, 16] have been extensively explored for functional prediction by detecting these clues from sequences. However, these methods tend to become less effective for low and non- homologous sequences [15, 17, 18]. In a comprehensive evaluation of sequence alignment methods against 15,208 enzymes labeled with an International Enzyme Commission EC class index, it has been found that approximately 60% of the EC classes containing two or more enzymes could not be perfectly discriminated by sequence similarity at any threshold [19]. The low and non-homologous proteins of unknown function constitute a substantial percentage, up to 20%~100%, of the open reading frames (ORFs) in many of the currently completed genomes [20]. Hence, it is highly desirable to explore methods less dependent of sequence similarity [3, 21, 22].

Progress has been made in developing alternative functional prediction methods to reduce the dependence on sequence similarity and clustering. Additional properties such as structural features [23, 24], interaction and network profiles [2, 6, 7, 25], and protein/gene fusion data [26, 27] have been used for predicting protein functions. More recently, machine learning methods have been explored for functional prediction from sequence-derived structural and physicochemical properties irrespective of sequence similarity [22, 28-38]. In particular, support vector machines (SVM) have shown promising potential for predicting proteins and peptides of various biochemical classes such as receptors [36, 39, 40], nucleic acid or lipid binding proteins [28, 38, 41, 42], and enzymes [30, 37, 43]), various therapeutic classes such as druggable proteins [44], hormone proteins [35], stress response proteins [35], cytokines [45], cytokine receptors [46], MHC-binding peptides [47], and other broadly defined functional classes such as crystallizable proteins [48], mitochondrial proteins [49], mesophilic and thermophilic proteins [50], and functional classes in yeast [51]).

This article analyzes the strategies, reported prediction performances, current progresses and difficulties in applying SVM for predicting various functional classes of proteins and peptides. Algorithms for representing proteins and peptides by using amino acid sequence derived structural and physicochemical descriptors [2, 31, 34, 52] are also discussed. A number of web-servers for facilitating the computation of these descriptors and for predicting the functional classes of proteins and peptides are also described.

FUNCTIONAL CLASSES OF PROTEINS AND PEPTIDES

Proteins have been classified into functional classes as well as sequence (domain) and structural families. The common functionality of class members primarily arise from the common structural and physicochemical properties at the active sites, which can be explored for predicting the function of proteins and peptides from sequence-derived struc-

*Address correspondence to this author at the Bioinformatics and Drug Design Group, Department of Pharmacy and Center for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543 and Shanghai Center for Bioinformatics Technology, Shanghai, 201203, P.R. China; Tel: +65-6874-6877; Fax: +65-6774-6756; E-mail: phacyz@nus.edu.sg

1389-2037/08 $55.00+.00

tural and physicochemical descriptors irrespective of sequence homology. This can be further demonstrated by evaluating enzyme, transporter, DNA-binding, and lipid-binding families. Enzymes represent the largest and most diverse group of all proteins, catalyzing chemical reactions in all organisms. Based on their catalyzed chemical reactions, enzymes can be divided into three levels of functional classes. The first level is composed of 6 super families (EC1 oxidoreductases, EC2 transferases, EC3 hydrolases, EC4 lyases, EC5 isomerases, and EC 6 ligases), the second level consists of 63 families (such as EC4.1 carbon-carbon lyases), and the third level contains 254 subfamilies (such as EC2.7.1 phosphotransferases with an alcohol group as acceptor). Active sites of enzymes are inherently reactive environments packed with specific types of amino acid residues and cofactors, and these and other structural features facilitate binding and catalysis of specific types of substrates [30].

The second example is transporters that play key roles in transporting cellular molecules across cell and cellular compartment boundaries, mediating the absorption and removal of various molecules, and regulating the concentration of metabolites and ionic species [53-55]. Specific transporters have been explored as therapeutic targets [56-58] and a variety of transporters are responsible for the absorption, distribution and excretion of drugs [59, 60]. Thus functional assignment of transporters is important for facilitating drug discovery and investigation of cellular processes and diseases. There are active and passive transporters. Active transporters couple solute transport to the input of energy and these can be divided into two classes: ion-coupled and ATP-dependent transporters. Ion-coupled transporters link uphill solute transport to downhill electrochemical ion gradients. ATP-dependent transporters are directly energized by the hydrolysis of ATP and they transport a heterogeneous set of substrates. Passive transporters include facilitated transporters and channels, which allow the diffusion of solutes across membranes. These transporters evolve from common themes into families of different architectures [53, 61, 62]. Transporters are divided into TC families based on their mode of transport, energy coupling mechanism, molecular phylogeny and substrate specificity [62]. TC families are classified at four levels (TC class, TC sub-class, TC family, and TC sub-family) as indicated by a specific TC number TC I.X.J.K.L. Here I=1, …, 9 represents each of the 9 TC classes, X=A, B, C, D, E, … represents each of the TC sub-classes that belong to a TC class, J=1, … represents each of the TC families that belong to a TC sub-class, K=1, … represents each of the TC sub-families that belong to a TC family, and L=1, … represents individual transporters under a sub-family.

The third example is DNA binding proteins that play critical roles in regulating gene transcription, DNA replication, DNA packaging, and DNA repair [63]. Prediction of DNA-binding proteins is important for studying proteins involved in genetic regulation [64-66]. DNA recognition by proteins is primarily mediated by combination of such structural and physicochemical features as specific DNA binding domains [67, 68], helix structures [67], minor groove binding architectures [68], asymmetric phosphate charge neutralization [68], conserved amino acids [69], hydrogen bonds [70], water-mediated bonds [70, 71], and indirect recognition mechanism [72]. DNA-binding proteins can be further divided into 9 major functional classes plus several smaller ones (such as covalent protein-DNA linkage proteins and terminal addition proteins). The 9 major classes are DNA condensation (for wrapping DNA around histones), DNA integration (mediating the insertion of duplex DNA into a chromosome), DNA recombination (for cleaving and rejoining DNA), DNA repair, DNA replication, DNA-directed DNA polymerase (catalyzing DNA synthesis by adding deoxyribonucleotide units to a DNA chain using DNA as a template), DNA-directed RNA polymerase (catalyzing RNA synthesis by adding ribonucleotide units to a RNA chain using DNA as a template), repressor (interfering with transcription by binding to specific sites on DNA), and transcription factor.

The fourth example is lipid-binding proteins that play important roles in cell signaling and membrane trafficking [73], lipid metabolism and transport [74, 75], innate immune response to bacterial infections [76], and regulation of gene expression and cell growth [77]. Prediction of the functional roles of lipid-binding proteins is important for facilitating the study of various biological processes and the search of new therapeutic targets. Lipid-binding proteins are diverse in sequence, structure, and function [78-85]. Non-the-less, lipid recognition by proteins is primarily mediated by some combination of a number of structural and physicochemical features including conserved fold elements [77], specific lipid-binding site architectures [78] and recognition motifs [79, 85], ordered hydrophobic and polar contacts between lipid and protein [80], and multiple noncovalent interactions from protein residues to lipid head groups and hydrophobic tails [85]. There are 8 major lipid-binding classes, which are lipid degradation, lipid metabolism, lipid synthesis, lipid transport, lipid-binding, lipopolysaccharide biosynthesis, lipoprotein (proteins posttranslationally modified by the attachment of at least one lipid or fatty acid, e.g. farnesyl, palmitate and myristate), lipoyl (proteins containing at least one lipoyl-binding domain)

A typical example of functional peptide classes is the classes of MHC-binding peptides [47]. Peptide binding to MHC is critical for antigen recognition by T-cells. One of the mechanisms of immune response to foreign or self protein antigens is the activation of T-cells by the recognition of T-cell receptors of specific peptides degraded from these foreign or self proteins and transported to the surface of antigen presenting cells [86]. Peptides recognized by T-cells are potential tools for diagnosis and vaccines for immunotherapy of infectious, autoimmune, and cancer diseases [87]. In many respects, MHC-binding and other protein-binding peptides possess similar characteristics as proteins of specific functional classes in that they also share some structural and physicochemical features to facilitate the common function: binding to MHC or other proteins [88-90].

SUPPORT VECTOR MACHINE APPROACH FOR PREDICTING FUNCTIONAL CLASSES OF PROTEINS AND PEPTIDES

Support vector machines can be explored for predicting the function of proteins and peptides by determining whether their sequence-derived properties conform to those of known

proteins and peptides of a specific functional class [28, 30, 33, 43, 51]. The advantage of this approach is that more generalized sequence-independent characteristics can be extracted from the sequence derived structural and physicochemical properties of the multiple samples that share common functional or interaction profiles irrespective of sequence similarity, which can be used to derive classifiers [1, 2, 28, 30, 33, 43, 47, 51, 91-94] for predicting other proteins and peptides that have the same functional or interaction profiles.

The task of predicting the functional class of a protein or peptide can be considered as a two-class (positive class and negative class) classification problem for separating members (positive class) and non-members (negative class) of a functional or interaction class. SVM and other well established two-class classification-based machine learning methods can then be applied for developing an artificial intelligence system to classify a new protein or peptide into the member or non-member class, which is predicted to have a functional or interaction profile if it is classified as a member. Sequence-derived structural and physicochemical properties have frequently been used for representing proteins and peptides [1, 2, 28, 30, 33, 43, 47, 51, 91-93] in the development of SVM and other machine learning classification systems for predicting the functional and interaction profiles of proteins.

Fig. **1** illustrates the process of using SVM for training and predicting proteins or peptides that have a specific common functional or interaction profile. Proteins or peptides known to have and not have the profile are represented by separate sets of feature vectors, which are composed of descriptors derived from the sequence of these proteins or peptides for representing their structural and physicochemical properties. These two sets of feature vectors are projected into a multi-dimensional space in which they are separated by a hyper-plane in such a way that those having the profile are on one side and those without the profile are on the other side of the hyper-plane. A new protein or peptide can be predicted to have the same profile if its feature vector is projected on the side of the hyper-plane where other proteins or peptides having the profile are located.

REPRESENTATION OF PROTEIN AND PEPTIDE SEQUENCES

Protein or peptide sequences have been represented by a number of amino acid sequence derived structural and physicochemical descriptors [2, 31, 34, 52]. These include amino acid composition, dipeptide composition, sequence autocorrelation descriptors, sequence coupling descriptors, and the descriptors for the composition, transition and distribution of hydrophibicity, polarity, polarizibility, charge, secondary structures, and normalized Van der Waals volumes. Web servers such as PROFEAT [95] (http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi) and ProtParam [96] (http://www.expasy.org/tools/protparam.html) have appeared for facilitating the computation of these descriptors. CBS Prediction Servers (http://www.cbs.dtu.dk/services/) can be used for computing other sequence derived features such as cleavage sites, nuclear export signals, and subcellular localization.

Amino acid composition is the fraction of each amino acid type in a sequence $f(r) = N_r / N$, where r=1,2,3, …, 20, N_r is the number of amino acid of type r and N is sequence length. Dipeptide composition is defined as $fr(r,s) = N_{rs} /(N-1)$, where r,s=1,2,3,…,20, and N_{ij} is the number of dipeptide represented by amino acid type r and s [36]. Autocorrelation descriptors are defined from the distribution of amino acid properties along the sequence [97]. The amino acid indices used in these auto-correlation descriptors include hydrophobicity scales [98],average flexibility indices [99], polarizability parameter [100], free energy of solution in water [100], residue accessible surface area in trepeptide [101], residue volume [102], steric parameter [103], and relative mutability [104]. Each of these indices is centralized and standardized before the calculation. Moreau-Broto autocorrelation descriptors are defined as $AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d}$, where d is the lag of the autocorrelation, P_i and P_{i+d} are the properties of the amino acids at position i and i+d respectively. The normalized Moreau-Broto autocorrelation descriptors are defined as $ATS(d) = AC(d)/(N-d)$, where d=1, 2, 3, … , 30. Moran autocorrelation descriptors are defined as:

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \overline{P})(P_{i+d} - \overline{P})}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \overline{P})^2} \quad \text{d=1, 2, 3, …, 30.}$$

where P_i and P_{i+d} are defined in the same way as above, and $\overline{P}$ is the average of the considered property P along the sequence. Geary autocorrelation descriptors are defined as:

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \overline{P})^2} \quad \text{d=1, 2, 3, …, 30.}$$

where $\overline{P}$, P_i and P_{i+d} are defined in the same way as in the above.

The quasi-sequence-order descriptors are derived from both the Schneider-Wrede physicochemical distance matrix [105-107] and the Grantham chemical distance matrix [108] between the 20 amino acids. The d-th-rank sequence-order-coupling number is defined as $\tau_d = \sum_{i=1}^{N-d}(d_{i,i+d})^2$ d=1, 2, …, 30 , where $d_{i,i+d}$ is the distance between the two amino acids at position i and i+d. For each amino acid type, the type-1 quasi-sequence-order descriptor can be defined as:

$$Xr = \frac{f_r}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{30}\tau_d} \quad \text{r=1, 2, 3, … , 20}$$

where f_r is the normalized occurrence for amino acid type i and w is a weighting factor(w=0.1). The type-2 quasi-sequence-order is defined as:

$$Xd = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{30}\tau_d} \quad \text{d=21, 22, 23, …, 50}$$

Three descriptors, composition (C), transition (T) and distribution (D), are derived for each of the following physicochemical properties: hydrophibicity, polarity, polarizibility, charge, secondary structures, and normalized Van der Waals volume [34, 109, 110]. For each of these properties, the constituent amino acids in a protein or peptide are divided in three classes according to its attribute such that each amino acid is encoded by one of the indices 1, 2, 3 according to which class it belongs to. For instance, amino acids can be divided into hydrophobic (CVLIMFW), neutral (GASTPHY), and polar (RKEDQN) groups. C is the number of amino acids of a particular property (such as hydrophobicity) divided by the total number of amino acids in a protein sequence. $C = (\frac{n_1 \times 100}{N}, \frac{n_2 \times 100}{N}, ..., \frac{n_m \times 100}{N})$ and $N = \sum_{i=1}^{i=m} n_i$. Here m=3 represents that 20 amino acids are divided into 3 property groups and n_i is the number of amino acid of a particular property. T characterizes the percent frequency with which amino acids of a particular property is followed by amino acids of a different property. $T = (\frac{T_{G1G2} * 100}{N-1}, \frac{T_{G1G3} * 100}{N-1}, \frac{T_{G2G3} * 100}{N-1})$ T_{GiGj} is the number of occurrence of amino acid of property i is followed by amino acid of property j. N-1 is the total number of transition within this protein sequence. D measures the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property is located respectively. $D = (D1, D2, D3)$, and $Di = (\frac{P_{i0} * 100}{N}, \frac{P_{i25} * 100}{N}, \frac{P_{i50} * 100}{N}, \frac{P_{i75} * 100}{N}, \frac{P_{i100} * 100}{N})$. P_{i0} and P_{i100} respectively represent the location of the first and last amino acids of property i in the protein sequence. Overall, there are 21 elements representing these three descriptors: 3 for C, 3 for T and 15 for D.

Construction of protein or peptide feature vectors can be illustrated by the generation of the amino acid composition descriptors of a hypothetical peptide (AEAELEAAEEAEMEAAE). This sequence contains 7 alanines (n1=7), 7 glutamic acids (n2=8), 1 leucine and 1 methionoine (n3=2).The composition is $C = (\frac{n1 \times 100}{n1+n2+n3}, \frac{n2 \times 100}{n1+n2+n3}, \frac{n3 \times 100}{n1+n2+n3}) = (41.18, 47.06, 11.76)$. There are 9 A=>E and E=>A transitions,2 E=>L and L=>E transitions, 2 E=>M and M=>E and 0 A=>M and M=>A transitions in this sequence. The total number of transitions is 16. The percent frequency of transition is thus $T = (T1,T2,T3) = (\frac{9}{16} \times 100, \frac{4}{16} \times 100, \frac{0}{16} \times 100) = (56.25, 0.25, 0)$. The first, 25%, 50%, 75% and 100% of As, Es and Ms with Ls is located within the 1st, 3rd, 8th, 11th, 16th residues, the 2nd, 4th, 9th, 10th, 17th residues and 5th, 5th , 5th , 13th,13th respectively. The distribution is then

$$D = (\frac{1}{17} \times 100, \frac{3}{17} \times 100, \frac{8}{17} \times 100, \frac{11}{17} \times 100, \frac{16}{17} \times 100, \frac{2}{17} \times 100, \frac{4}{17} \times 100, \frac{9}{17} \times 100, \frac{10}{17} \times 100, \frac{17}{17} \times 100, \frac{5}{17} \times 100, \frac{5}{17} \times 100, \frac{5}{17} \times 100, \frac{13}{17} \times 100, \frac{13}{17} \times 100)$$
$$= (5.88, 17.65, 47.06, 67.71, 94.12, 11.76, 23.53, 52.94, 58.82, 100.00, 29.41, 29.41, 29.41, 76.47, 76.47)$$

Overall, the amino acid composition feature vector is

X=(C, T, D)=(41.18, 47.06, 11.76, 56.25,0.25,0, 5.88, 17.65, 47.06, 67.71, 94.12, 11.76, 23.53, 52.94, 58.82, 100.00,29.41, 29.41, 29.41,76.47, 76.47)

for this sequence. All generated vectors have equal length, which is useful for classification of proteins and peptides of variable lengths by using statistical learning methods.

ALGORITHMS AND SOFTWARE TOOLS OF SUPPORT VECTOR MACHINES

There are linear and nonlinear SVM algorithms. Linear SVM directly constructs a hyperplane in the feature space to separate positive examples from negative examples. On the other hand, nonlinear SVM projects both positive and negative examples into a higher-dimensional feature space and then separates them in that space. The following is a brief description of the algorithms of linear and nonlinear SVM. SVM software tools and SVM-based servers for predicting functional class of proteins and peptides are listed in Table **1**.

Linear SVM

Let the training data of two separate classes, each containing *n* samples, be represented by $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$,…, $(\mathbf{x}_n, y_n)$, $i = 1, 2,\ldots, n$, where $\mathbf{x}_i \in R^N$ is a vector in an *N*-dimensional space representing various physicochemical and structural properties of a protein or peptide, and $y_i \in (-1, +1)$ indicates class label (e.g. (+) represents members and (-) non-members of a functional class). Given a weight vector **w** and a bias *b*, it is assumed that these two classes can be separated by two margins parallel to the hyper-plane as illustrated in (Fig. **2a**), which can be represented as a single inequality:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \text{ for } i = 1, 2, \ldots, n \qquad (1)$$

where $\mathbf{w} = (w_1, w_2,\ldots, w_n)^T$ is a vector of *n* elements. As shown in (Fig. **2b**), there are a number of separate hyper-planes for an identical group of training data. The objective of SVM is to determine the optimal weight w_0 and optimal bias b_0 such that the corresponding hyper-plane separates S+ and S- with a maximum margin and gives the best prediction performance. This hyper-plane is called Optimal Separating Hyper-plane (OSH) as illustrated in (Fig. **2c**).

The equation for a hyper-plane can be written as:

$$\mathbf{w} \cdot \mathbf{x}_i + b = 0 \qquad (2)$$

and the distance between the two corresponding margins is:

Table 1. Web-servers for Computing Functional Class of Proteins and Peptides by Using Support Vector Machines. Web-Sites of Support Vector Machine Software are also given

Category	Web-server or Software	URL
Server for Predicting Protein Functional Class	CTKPred: SVM prediction and classification of the cytokine family	http://bioinfo.tsinghua.edu.cn/~huangni/CTKPred/
	GPCRpred: SVM prediction of families and subfamilies of G-protein coupled receptors	http://www.imtech.res.in/raghava/gpcrpred/info.html
	pSLIP: SVM protein subcellular localization prediction	http://pslip.bii.a-star.edu.sg/
	SVMProt: SVM protein functional family prediction from protein sequence	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi
Server for Predicting Peptide Functional Class	MHC-BPS: SVM prediction of MHC-binding peptides of flexible lengths	http://bidd.cz3.nus.edu.sg/mhc/
	SVMHC: SVM prediction of MHC-binding peptides.	http://www.sbc.su.se/svmhc/
	SVRMHC: SVM prediction of MHC binding peptide	http://svrmhc.umn.edu/SVRMHCdb/
	WAPP: SVM prediction of MHC-binding, proteasomal cleavage and TAP transport peptides	http://www-bs.informatik.uni-tuebingen.de/WAPP
SVM Software and servers	BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/
	GIST SVM Server	http://svm.nbcr.net/cgi-bin/nph-SVMsubmit.cgi
	LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
	LS-SVMlab	http://www.esat.kuleuven.ac.be/sista/lssvmlab/
	mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
	SVM light	http://svmlight.joachims.org/
	SMO	http://www.datalab.uci.edu/people/xge/svm/
	WinSVM	http://www.cs.ucl.ac.uk/staff/M.Sewell/winsvm/

$$\gamma(\mathbf{w},b) = \min_{\{\mathbf{x}|y=+1\}} \frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|} - \max_{\{\mathbf{x}|y=-1\}} \frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|} \quad (3)$$

The OSH can be obtained by maximizing the above distance or minimizing the norm of ||**w**|| under inequality constraints (Eq. (1)), and

$$\gamma_{\max} = \gamma(\mathbf{w}_0, b_0) = \frac{2}{\|\mathbf{w}\|} \quad (4)$$

The saddle point of the following Lagrangian gives solutions to the above optimization problem:

$$L(\mathbf{w},b,\alpha) = \frac{1}{2}\mathbf{w}\cdot\mathbf{w} - \sum_{i=1}^{n}\alpha_i\left[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1\right] \quad (5)$$

where $\alpha_i \geq 0$ are Lagrange multipliers. The solution to this optimization Quadratic Programming (QP) problem requires that the gradient of $L(\mathbf{w}, b, \alpha)$ with respect to **w** and b vanishes, i.e. $\partial L/\partial \mathbf{w}|_{\mathbf{w}=\mathbf{w}_0} = 0$ and $\partial L/\partial b|_{b=b_0} = 0$, which gives rise to the following conditions:

$$\mathbf{w}_0 = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \quad (6)$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \quad (7)$$

By substituting Eqs. (6) and (7) into Eq. (5), the QP problem becomes the maximization of the following expression:

$$L(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j) \quad (8)$$

under the constraints

$$\sum_{i=1}^{n}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, n \quad (9)$$

where C is a penalty for training errors for soft-margin SVM and is equal to infinity for hard-margin SVM.

The points located on the two optimal margins will have nonzero coefficients α_i among the solutions to Eq. (8), and

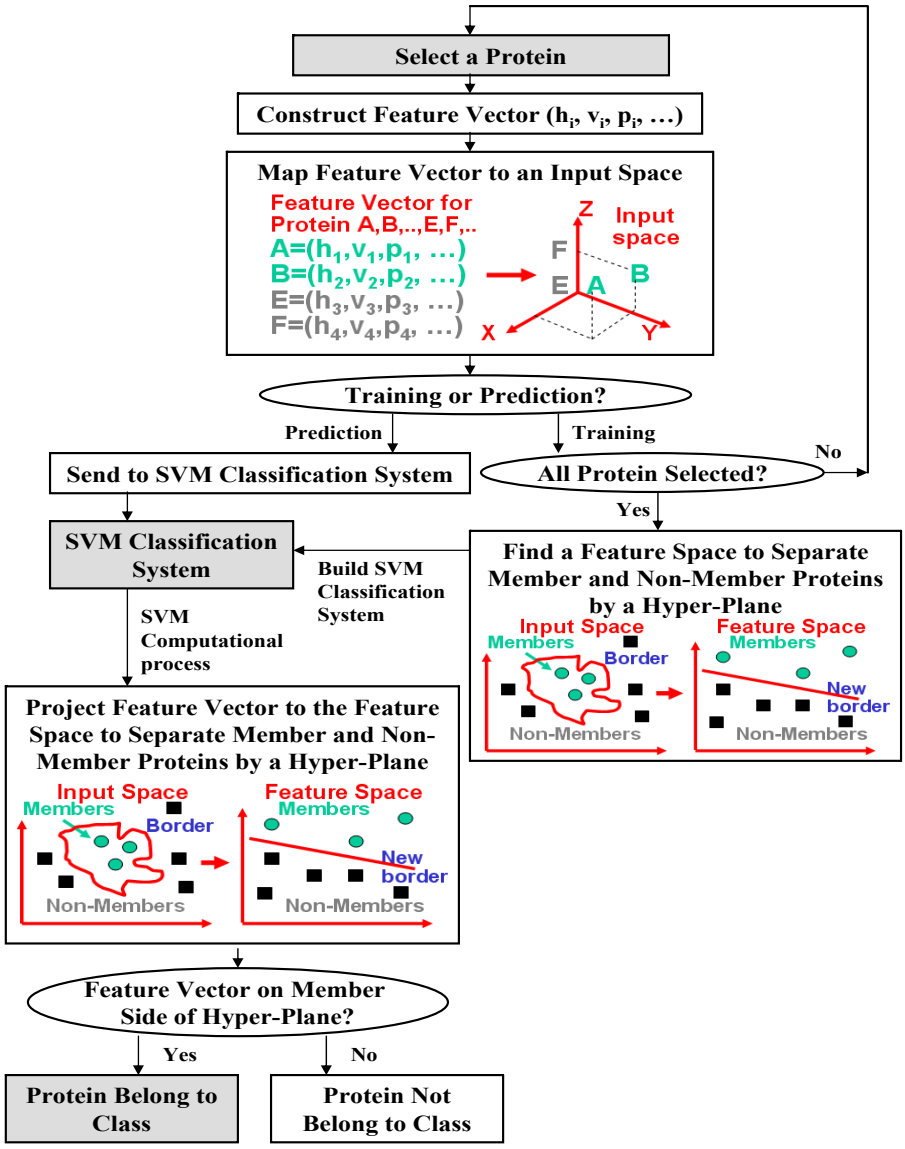

Fig. (1). Schematic diagram illustrating the process of the training and prediction of the functional class of proteins and peptides by using support vector machine (SVM) method. A,B: feature vectors of proteins belong to a functional class; E,F: feature vectors of proteins not belong to a functional class. Sequence-derived feature h_j, p_j, vj … represents such structural and physicochemical properties as hydrophobicity, polarizability, and volume; or such properties as domain information, subcellular localization, and post-translational (PT) modification profiles etc.

are called *Support Vectors* (SV). The bias b_0 can be calculated as follows:

$$b_0 = -\frac{1}{2}\left\{\min_{\{\mathbf{x}_i | y_i = +1\}} (\mathbf{w}_0 \cdot \mathbf{x}_i) + \max_{\{\mathbf{x}_i | y_i = -1\}} (\mathbf{w}_0 \cdot \mathbf{x}_i)\right\} \quad (10)$$

After determination of support vectors and bias, the decision function that separates the two classes can be written as:

$$f(\mathbf{x}) = \text{sign}\left[\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b_0\right] = \text{sign}\left[\sum_{SV} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b_0\right] \quad (11)$$

Nonlinear SVM

Real-world problems are usually nonlinear in nature. The linear classification scheme described above thus becomes unsuitable to these problems. A nonlinear classification scheme can be introduced such that the original training data **x** in the input space **X** is projected into a high-dimensional feature space **F** via a Mercer kernel operator *K* [111], followed by the construction of OSH in the feature space as illustrated in (Fig. **2d**). In other words, the set of classifiers is transformed into the form:

$$f(\mathbf{x}) = \text{sign}\left[\sum_{SV} \alpha_i y_i K(\mathbf{x}_i \cdot \mathbf{x}) + b_0\right] \quad (12)$$

where *K* is a symmetric positive definite function that satisfies Mercer's conditions:

$$K(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^{\infty} \alpha_m \phi(\mathbf{x}) \cdot \phi(\mathbf{y}), \ \alpha_m \geq 0,$$
$$\iint K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) dx dy > 0, \ \int g^2(\mathbf{x}) dx < \infty \quad (13)$$

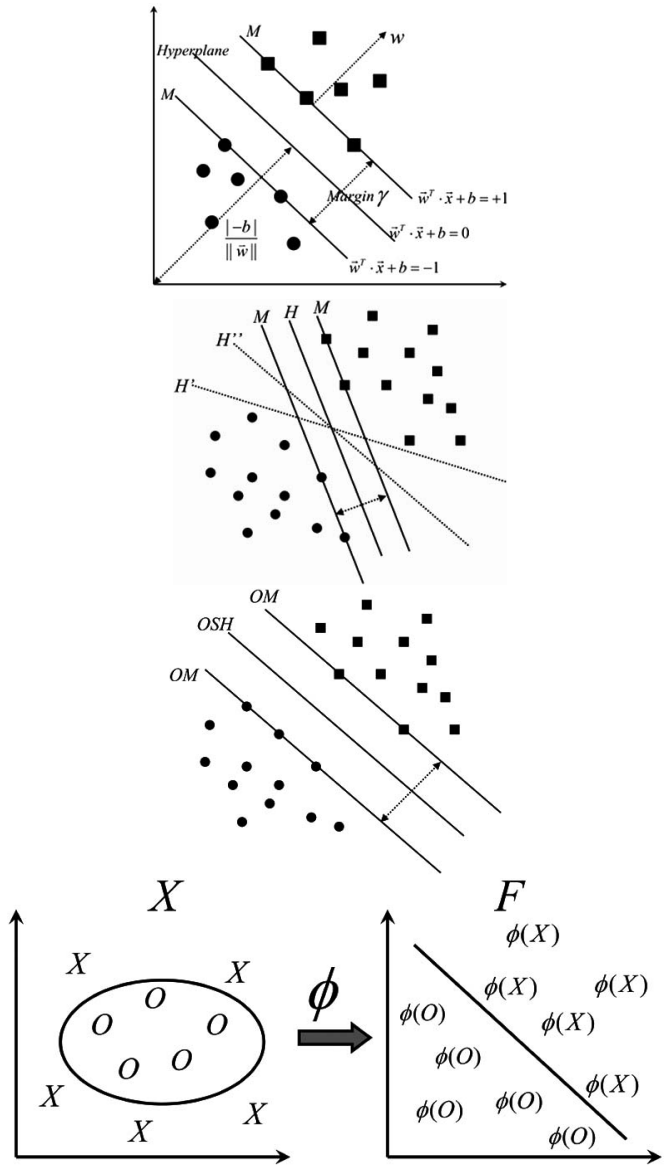

Fig. (2) Support vector machines. (a) Definition of hyper-plane and margin. The circular dots and square dots represent samples of class -1 and class +1, respectively. (b) The available hyper-planes *H*, *H'*, *H''*,…, corresponding to a set of training data. (c) Unique optimal separating hyper-plane of a set of training data. (d) Basic idea of support vector machines: Projection of the training data nonlinearly into a higher-dimensional feature space via ϕ, and subsequent construction of a separating hyper-plane with maximum margin in that space.

in which the Kernel represents a legitimate inner product in the input space:

$$K(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y}) \tag{14}$$

In the feature **F**-space, the dual Lagrangian, given in Eq. (8), is:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i \cdot \mathbf{x}_j) \tag{15}$$

subject to the same constraints (Eq. (9)) as Eq. (8). Thus the decision function is now given by Eq. (12) in which

$$b_0 = -\frac{1}{2}\left\{ \min_{\{\mathbf{x}_i | y_i = +1\}} \left(\sum_{\mathrm{SV}} \alpha_i y_i K(\mathbf{x}_i \cdot \mathbf{x}_j) \right) + \max_{\{\mathbf{x}_i | y_i = -1\}} \left(\sum_{\mathrm{SV}} \alpha_i y_i K(\mathbf{x}_i \cdot \mathbf{x}_j) \right) \right\} \tag{16}$$

A number of kernel functions have be used in SVM. Examples of the most popular ones are:

Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i \cdot \mathbf{x}_j + 1\right)^p$

Gaussian: $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\left\|\mathbf{x}_j - \mathbf{x}_i\right\|^2 / 2\sigma^2}$

Sigmoidal: $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\left\|\mathbf{x}_j - \mathbf{x}_i\right\|^2 / 2\sigma^2}$ (17)

as well as their combinations in the form of summation or tensor products of these kernels. A vector has a limited number of components, each representing a specific physicochemical, structural or biological quantity. Each quantity is normalized or scaled, such that its value is of finite value. From a practical point of view, $\mathbf{x} \cdot \mathbf{y}$ is of finite value so as to avoid the value of polynomial kernel reaching infinity.

METHODS FOR TRAINING, TESTING AND ESTIMATING GENERALIZATION CAPABILITIES OF SUPPORT VECTOR MACHINES CLASSIFICATION SYSTEMS

Several validation methods have been used for training, testing, and estimating generalization errors of a SVM model [36, 92, 112, 113] based on a "re-sampling" strategy [114, 115]. The commonly used validation methods include N-fold cross validation, leave one out, leave v out, jack-knifing, and bootstrapping. In N-fold cross validation, samples are randomly divided into N subsets of approximately equal size. N-1 subsets are used as a training set for developing a SVM model, and the remaining one is used as a testing set for evaluating the prediction performance of that model. This process is repeated N times such that every subset is used as a testing set once. The average accuracy of the N number of SVM models is used for measuring the generalization capability of the SVM method. When N equals to the total number of samples, the method is called “leave one out” such that every sample is used for testing a SVM model trained by using all of the other samples. "Leave-v-out" is a more elaborate and expensive version of the “leave something out” cross-validation that involves leaving out all possible combinations of v samples as a test set. In jack-knifing, samples are distributed and used for training and testing the SVM models in the same way as that of “leave one out” method, but the generalization error of the derived SVM models is estimated based on the comparison of the average accuracy of subsets and that of all sets of these SVM models. In bootstrapping, different combinations of randomly selected subsets of samples are separately used for training SVM models each of which is tested by using the compounds not included in the respective training set.

Moreover, independent evaluation sets have also been used for testing the performance of SVM classification systems [34, 42, 116, 117]. In using this approach, samples are divided into training, testing, and independent validation set based on their distribution in protein or peptide descriptor space. Protein or peptide descriptor space is defined by the commonly used structural and chemical descriptors of proteins or peptides. Samples can be clustered into groups based on their distance in the descriptor space by using such methods as hierarchical clustering [118]. An upper-limit of the largest separation of r can be used for restricting the size of each cluster. One or more representative samples are randomly selected from each group to form a training set that is sufficiently diverse and broadly distributed in the chemical space. One or more of the remaining compounds in each group are randomly selected to form the testing set. The remaining samples are used as the independent evaluation set, which show reasonable level of structural diversity and distinction with respect to compounds of other groups.

The performance of SVM has been measured by using the positive prediction accuracy P_+ for proteins that have a specific property and the negative prediction accuracy P_- for proteins without that property [1, 2, 28, 30, 33, 43, 47, 51, 91-94]. Moreover, an overall accuracy P=(TP+TN)/N, where TP and TN is the true positive and true negative respectively and N is the number of proteins or peptiodes, can also be used to indicate the overall prediction performance. In some cases, P, P_+, and P_- are insufficient to provide a complete assessment of the performance of a discriminative method [119, 120]**.** Thus the Matthews correlation coefficient $MCC = (TP \times TN - FP \times FN) / \sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}$ has been used for measuring the performance of support vector machine [30, 33, 36, 39, 45, 49].

ASSESSMENT OF THE PERFORMANCE OF SUPPORT VECTOR MACHINE CLASSIFICATION SYSTEMS

Performance for Predicting Functional Classes of Proteins and Peptides

Table **2** summarizes the reported performance of the use of SVM for predicting protein functional classes. The reported P_+ and P_- values are in the range of 25.0%~100.0% and 69.0%~100.0%, with the majority concentrated in the range of 75%~95% and 80%~99.9% respectively. Based on these reported results, SVM generally shows certain level of capability for predicting the functional class of proteins and protein-protein interactions. In many of these reported studies, the prediction accuracy for the non-members appears to be better than that for the members. The higher prediction accuracy for non-members likely results from the availability of more diverse set of non-members than that of members, which enables SVM to perform a better statistical learning for recognition of non-members.

The performance of SVM for predicting functional classes of peptides are given in Table **3**. Prediction of protein-binding peptides have primarily been focused on MHC-binding peptides [47], the reported P_+ and P_- values for MHC binding peptides are in the range of 75.0% ~ 99.2% and 97.5% ~ 99.9%, with the majority concentrated in the range of 93.3%~95.0% and 99.7%~99.9% respectively. These studies have demonstrated that, apart from the prediction of protein functional classes, SVM is equally useful for predicting protein-binding peptides.

Performance for Predicting Functional Classes of Novel Proteins

The performance of SVM for predicting the functional profile of novel proteins has also been evaluated by several studies listed in Table **4**. These novel proteins are of two types. The first includes several groups of proteins that have no homologous counterpart in well-established protein database, and the second contains pairs of homologous enzymes that belong to different functional families. The non-homologous nature of the first type of novel proteins complicates the task of using sequence alignment and clustering methods for determining their functions. On the other hand, the homologous nature of the second type of novel proteins may result in false association of proteins of different functional families if sequence similarity is used as the sole indicator of functional association. Therefore, it is desirable to explore other methods with less or no reliance on homology to complement sequence similarity and clustering methods [3, 21] . From Table **4**, SVM appears to have the capacity of correct prediction of 46.3%~76.7% of the novel proteins found from the literatures.

Table 2. Performance of Machine Learning Methods for Predicting Functional Class of Proteins as Reported in the Literature. All of the Data and Results were Collected from the Original Papers. Please Refer to the Respective Reference for Complete Results. N+, N- and N are the Number of Class Members, Non-Members and All Proteins (Members + Non-Members) Respectively, P+ and P- are Prediction Accuracy for Class Members And Non-Members Respectively, P is the Overall Accuracy, and MCC is the Matthews Correlation Coefficient

Protein Functional Class	Protein Sub-Classes	Protein Descriptors	Number of Proteins in Training Set N (N+/N-)	Validation Method	Reported Prediction Accuracy				Ref
					P+ (%)	P- (%)	P (%)	MCC	
Enzymes	46 sub-classes: EC1.1~ EC1.11, EC1.13~EC1.15, EC1.17, EC1.18, EC2.1~EC2.8, EC3.1~EC3.6, EC4.1~EC4.4, EC4.6, EC5.1~EC5.5, EC5.99, EC6.1~EC6.5	Physicochemical properties	956~9216 (35~3892/ 807~5324)	Independent evaluation	53.0~ 99.3	85.0~ 99.7	81.8~99.7	0.31~ 0.98	[30, 34]
	54 sub-classes: EC1.1~EC1.21, EC2.1~EC2.8, EC3.1~EC3.8, EC4.1~EC4.6, EC5.1~Ec5.6, EC6.1~6.6	Functional Domain Composition and pseudo amino acid composition	503~3582 (3~2002/327~3548)	Jackknife	25.0~ 100.0				[37]
	39 pfam kinase families	Amino acid composition, dipeptide composition, amino acid co-occurence	N.A	Independent evaluation			95.7~95.9		[218]
	16 oxidoreductase subfamilies	Am-Pse-AAC descriptors	64~314	5-fold CV			50.8~96.8		[219]
	Glutathione S-transferases	Amino acid composition, dipeptide composition	214 (107/107)	5-fold CV			91.6~95.8		[220]
Transporters	20 sub-classes: TC1.A, TC1.A.1, TC1.B, TC1.E, TC2.A, TC2.A.1, TC2.A.3, TC2.A.6, TC2.C, TC3.A, TC3.A.1, TC3.A.3, TC3.A.5, TC3.A.15, TC3.D, TC3.E, TC4.A, TC8.A, TC9.A, TC9.B	Physicochemical properties	613~7508 (50~1220 /513~7299)	Independent evaluation	60.6~ 97.1	91.5~ 99.9	91.4~ 99.7	0.27~ 0.97	[221]

(Table 2) contd….

Protein Functional Class	Protein Sub-Classes	Protein Descriptors	Number of Proteins in Training Set N (N+/N-)	Validation Method	Reported Prediction Accuracy				Ref
					P+ (%)	P- (%)	P (%)	MCC	
	Voltage gated ion channels	Amino acid composition, dipeptide composition	236	5-fold CV			82.9~89.1	0.66~0.78	[222]
Cytokines	All cytokines	Dipeptide composition	1110 (437/673)	7-fold CV	92.5	97.2	95.3	0.90	[45]
	FGF/HBGF		437 (83/354)		92.7	98.6	97.5	0.92	
	TGF-β		437 (190/247)		97.4	94.7	95.8	0.92	
	TNF		437 (96/341)		94.0	98.8	97.7	0.94	
	Joint class (IL-6, LIF/OSM, MDK/PTN, NGF)		437 (68/369)		91.0	99.7	98.4	0.94	
	6 sub-classes: BMP, GDF, GDNF, INH, TGFB, other		N.A		46.7~ 100	85.5~ 100	84~ 98	0.65~ 0.96	
G-protein coupled receptors	All GPCRs	Physicochemical properties	2247 (927/1320)	Independent evaluation	95.6	98.1	97.4	0.93	[34]
		Dipeptide composition	3302 (778/2524)	5-fold CV	98.6	99.8	99.5	0.99	[39]
		Protein power spectrum	946	Jackknife			96.1		[38]
	Gi/o binding type	Structural characteristics (extra cellular loops, intracellular loops etc)	132 (61/71)	4-fold CV	77.0	78.3			[40]
	Gq/11 binding type		132 (47/85)	4-fold CV	68.1	72.7			
	Gs binding type		132 (24/108)	4-fold CV	83.3	95.2			
	Rhodopsin-like (Class A)	Protein power spectrum	540	Jackknife			97.0	0.93	[38]
		Amino acid composition	4350	Independent evaluation	82~100	28~100	73~99		[223]
	Secretin-like (Class B)	Protein power spectrum	187	Jackknife			96.3	0.94	[38]
	Metabotropic glutamate (Class C)	Protein power spectrum	103	Jackknife			94.2	0.95	[38]
	Fungal pheromone (Class D)	Protein power spectrum	21	Jackknife			81.0	0.92	[38]
	cAMP receptors (Class E)	Protein power spectrum	5	Jackknife			100	1.0	[38]

(Table 2) contd….

Protein Functional Class	Protein Sub-Classes	Protein Descriptors	Number of Proteins in Training Set N (N+/N-)	Validation Method	Reported Prediction Accuracy				Ref
					P+ (%)	P- (%)	P (%)	MCC	
	Friz-zled/smoothened (Class F)	Protein power spectrum	90	Jackknife			95.6	0.94	[38]
Cytokine recep-tors		Physicochemi-cal properties	1,002,243 (2,243/1000,000)	Independent evaluation	97.4	99.2		0.89	[46]
Nuclear recep-tors	All nuclear receptors	Amino acid composition	282	5-fold CV			82.6	0.74	[36]
		Dipeptide composition	282	5-fold CV			97.5	0.96	
		Physicochemi-cal properties	872 (334/538)	Independent evaluation	89.5	97.6			[34]
		Protein power spectrum	465	Jackknife			95.3		[38]
	Thyroid hor-mone-like	Protein power spectrum	165	Jackknife			95.8	0.95	[38]
	HNF4-like		114	Jackknife			97.4	0.96	
	Estrogen-like		130	Jackknife			97.7	0.96	
	Fushitarazu-F1 like		35	Jackknife			94.3	0.97	
	Nerve growth factor IB-like		5	Jackknife			80.0	0.89	
	Germ cell nuclear receptor		2	Jackknife			100	1.0	
	0A Knirps-like		7	Jackknife			42.9	0.65	
	0B DAX-like		7	Jackknife			71.4	0.84	
DNA/RNA-binding proteins		Amino acid periodicities of hysicochemical properties	4870	10-fold CV	58.2~72.7	79.8~87.5	76.0~81.5		[224]
RNA-binding proteins	All RNA-binding proteins	Amino acid composition and limited range correla-tion of hydrophobicity and solvent accessible surface area	6264 (1496/4768)	10-fold CV	76.5	97.2	92.2		[28]
		Physicochemi-cal properties	5126 (2161/2965)	Independent evaluation	97.8	96.0	96.1	0.80	[33]

(Table 2) contd….

Protein Functional Class	Protein Sub-Classes	Protein Descriptors	Number of Proteins in Training Set N (N+/N-)	Validation Method	Reported Prediction Accuracy				Ref
					P+ (%)	P- (%)	P (%)	MCC	
	rRNA-binding	Amino acid composition, limited range correlation of hydrophobicit, solvent accessible surface area	5824 (1056/4768)	10-fold CV	100	99.9	99.9		[28]
		Physicochemical properties	1680 (708/972)	Independent evaluation	94.1	98.7	98.6	0.74	[33]
	tRNA-binding	Physicochemical properties	886 (94/792)	Independent evaluation	94.1	99.9	99.8	0.92	[33]
	mRNA-binding		2383 (277/2106)		79.3	96.5	96.0	0.53	
	snRNA-binding		2021 (33/1988)		45.0	99.7	99.5	0.38	
DNA-binding proteins	All DNA-binding proteins	Amino acid composition, limited range correlation of hydrophobicity, solvent accessible surface area	12507 (7739/4768)	10-fold CV	92.8	77.1	86.8		[28]
		Surface and overall composition, overall charge and positive potential patches on the protein surface	359 (121/238)	5-fold CV	89.1	82.1	93.9		[41]
				Jackknife	90.5	81.8	94.9		
				leave 1-pair holdout	86.3	80.6	87.5		
				Leave-half holdout	83.3	82.5	83.5		
		Physicochemical properties	8575 (4240/4335)	Independent evaluation	90.9	87.6	88.5	0.74	[34, 225]
		Residue features (identity, charge, solvent accessibility, average potential, secondary structure, neighboring residues, location in a cationic patch)	50	Independent evaluation			78		[226]
		Physicochemical properties , pseudo amino acid composition, dipeptide composition	349 (118/231)	Jackknife	90.7	99.6	96.6	0.92	[227]

(Table 2) contd….

Protein Functional Class	Protein Sub-Classes	Protein Descriptors	Number of Proteins in Training Set N (N+/N-)	Validation Method	Reported Prediction Accuracy				Ref
					P+ (%)	P- (%)	P (%)	MCC	
	DNA condensation	Physicochemical properties	2410 (50/2360)	Independent evaluation	94.9	98.3	98.3	0.47	[34, 225]
	DNA integration		1307 (134/1173)		87.9	99.9	99.7	0.91	
	DNA recombination		3357 (889/2468)		87.8	98.9	97.9	0.87	
	DNA repair		5785 (2142/3643)		88.7	96.8	95.3	0.84	
	DNA replication		3734 (1131/2603)		85.6	96.6	95.4	0.79	
	DNA-directed DNA polymerase		2348 (273/2075)		72.9	99.7	98.9	0.79	
	DNA-directed RNA polymerase		2594 (484/2110)		90.8	99.4	98.8	0.91	
	Repressor		3684 (1337/2347)		93.3	95.6	95.4	0.76	
	Transcription factors		2354 (670/1684)		86.1	99.5	99.3	0.79	
Lipid-binding proteins	All lipid-binding proteins	Physicochemical properties	6933 (3232/3701)	Independent evaluation	89.9	97.0	94.1	0.88	[34, 42]
	Lipid transport		2262 (153/2109)		79.5	99.8	99.6	0.80	
	Lipid metabolism		2262 (293/1969)		79.5	99.2	98.8	0.72	
	Lipid synthesis		3498 (891/2607)		82.2	99.6	98.1	0.87	
	Lipid degradation		2178 (403/1775)		78.9	99.9	99.3	0.87	
Metal-binding proteins	All metal-binding	Physicochemical properties	33295	Independent evaluation	78.1	80.0	80.6		[228]
	Calcium-binding		5426		86.3	88.2	87.9		
	Cobalt-binding		1467		81.6	99.9	98.9		
	Copper-binding		1645		83.5	98.1	97.3		
	Iron-binding		9462		94.0	91.4	91.7		
	Magnesium-binding		9688		81.2	87.9	85.3		
	Manganese-binding		4214		85.4	94.5	93.2		

(Table 2) contd….

Protein Functional Class	Protein Sub-Classes	Protein Descriptors	Number of Proteins in Training Set N (N+/N-)	Validation Method	Reported Prediction Accuracy				Ref
					P+ (%)	P- (%)	P (%)	MCC	
	Nickel-binding		705		77.6	99.2	98.6		
	Potassium-binding		1240		90.4	99.9	99.6		
	Sodium-binding		1567		90.9	99.9	99.4		
	Zinc-binding		16072		74.9	98.0	86.7		
Transmembrane proteins		Functional Domain Composition	2059	jackknife test			86.3		[229]
				independent test			67.5		
				self-consistency			93.9		
		Pseudo-amino acid composition	2059	jackknife test			82.4		[230]
				independent test			90.3		
				self-consistency			99.9		
		Physicochemical properties	4668 (2105/2563)	Independent evaluation	90.1	86.7	86.7	0.75	[34]
Outer membrane proteins		Physicochemical properties	11,367 (1492/9872)	Independent evaluation	92.7	86.4			[34]
		Amino acid composition , dipeptide composition	1319 (377/942)	5-fold CV	63.5~88.0	93.8~100	91.1~94.4	0.75~0.84	[231]
Mitochondrial proteins		Amino acid composition	10372 (1432/8940)	5-fold CV	78.9	90.0	88.2	0.62	[191]
		Dipeptide composition	1180 (499/681)	Jackknife	79~80	85~89	83~85	0.65~0.69	[232]
Druggable proteins or drug targets		Physicochemical properties	28,355 (1484/26,871)	5-fold CV	67.6	83.6			[44]
		Physicochemical properties, autocorrelations	17,823 (1042/16781)	10-fold CV	72.3	98.5			[233, 234]
		Amino acid composition, physicochemical properties	6138 (186/5952)	10-fold CV	85	80	83		[235]
Allergenic proteins		Amino acid	1278 (578/700)	Independent evaluation	88.9	81.9	85.0	0.71	[236]
		Dipeptide composition	1278 (578/700)	Independent evaluation	82.8	85.0	84.0	0.68	
		Physicochemical properties	23474 (1005/22469)	Independent evaluation	93.0	99.9	99.7	0.96	[237]

(Table 2) contd….

Protein Functional Class	Protein Sub-Classes	Protein Descriptors	Number of Proteins in Training Set N (N+/N-)	Validation Method	Reported Prediction Accuracy				Ref
					P+ (%)	P- (%)	P (%)	MCC	
		Alignment scores against filtered length-adjusted allergen peptides	52843 (762/52081)	Independent evaluation	>75%				[238]
Crystallizable proteins		Mono-, di-, tripeptide composition, physicochemical and structural properties	923 (721/202)	10-fold CV	65.0	69.0	67.0		[48]
Mesophilic or thermophilic proteins		Amino acid composition	8417 (4895/3522)	5-fold CV and Independent evaluation			90.5~92.4		[239]
		Amino acid composition	5391 (3400/1991)	5-fold CV	82.2	92.9	89.2		[50]
Aggregating proteins		RQA descriptors	148 (39/109)	10-fold CV	86.3	70.8	84.0		[240]
Mostly-disordered proteins		RQA descriptors	N.A	10-fold CV	89.5	70.8	85.6		[240]
Malaria adhesins and adhesin-like proteins		Dipeptide composition, multiplet composition	5411	5-fold CV	100	99.9			[241]
Functional classes in yeast	All proteins	Functional domain composition	4902	Jackknife			72.0		[51]
	13 classes: Metabolism, energy, cell growth, cell division, DNA synthesis, transcription, protein synthesis, protein destination, transport facilitation, intracellar transport, cellular biogenesis, signal transduction, cell rescue, ionic homeostasis, cellular organization		86~725	Jackknife			15~ 90		

Table 3. Performance of Support Vector Machine Prediction of Functional Classes of Peptides. N+ and N- are the Number of Members and Non-Members in a Class, P+ and P- are the Reported Prediction Accuracy for Members and Non-Member Respectively, and P is the Reported Overall Accuracy

Peptide Class	Peptide Descriptors	Number of Peptides in Training Set N (N+/N-)	Validation Method [a] (N+/N-)	Reported Prediction Accuracy			Reference
				P_+ (%)	P_- (%)	P (%)	
Antibacterial peptides	Amino acid composition	872 (436/436)	Validation set (39/?)	89.9	88.1	89.0	[242]
A-Conotoxins	Polarity index attribute (residue surface burriability, polarity, and hydropathy)	176 (116/60)	Jackknife	85.0	95.5		[243]
M-Conotoxins				84.6	100.0		
O-Conotoxins				82.0	96.2		
T-Conotoxins				76.5	92.9		
Binders of HLA Allele A0201	Orthogonal factors from physical properties	203 (36/167)	10-fold CV	46.3~76.3	71.2~89.8	71.6~86.7	[244]
	Amino acid sequence	113	10-fold CV	90.0		78.0 (Mc)	[245]
	physicochemical properties	8036 (1125/6911)	Validation set (130/6664)	99.2	97.5	97.5	[246, 247]
Binders of HLA Allele A1	Amino acid sequence	28	10-fold CV	98.0		96.0(Mc)	[245]
	physicochemical properties	7031 (200/6831)	Validation set (40/6830)	75.0	99.7	99.6	[246, 247]
Binders of HLA Allele A2	Variant residues, binding environment	3050 (664/2386)	10-fold CV	90	90		[248, 249]
Binders of HLA Allele A3	Amino acid sequence	73	10-fold CV	91.0		80.0(Mc)	[245]
	physicochemical properties	6972 (139/6833)	Validation set (30/6833)	93.3	98.8	98.7	[246, 247]
	Variant residues, binding environment	2216 (680/1536)	10-fold CV	92	90		[248, 249]
Binders of HLA Allele B8	Amino acid sequence	25	10-fold CV	91.0		79.0(Mc)	[245]
	physicochemical properties	7001 (168/6833)	Validation set (20/6830)	95.0	99.8	99.8	[246, 247]
Binders of HLA Allele B2705	Amino acid sequence	29	10-fold CV	100.0		100.0(Mc)	[245]
	physicochemical properties	7502 (141/7361)	Validation set (21/7359)	95.0	99.9	99.9	[246, 247]
Binders of HLA Allele DR	Variant residues, binding environment	2396 (448/1948)	10-fold CV	>80			[248, 249]
Binders of HLA Allele DRB1.0401	Binary code of amino acid sequence	567	5-fold CV	80.2~87.1	77.4~85.0	78.8~86.1	[250]
	physicochemical properties	7422 (539/6883)	Validation set (100/6704)	95.0	99.9	99.9	[246, 247]
T-cell epitopes	Biobasis function		10-fold CV			90.3	[251]

The ability of SVM in predicting the functional profile of the first type of novel proteins have been attributed to the non-discriminative nature of SVM for selecting class members, and to the use of structural and physicochemical descriptors for representing proteins [20, 121-124]. In some cases, protein function is determined by specific structural and chemical features at active sites, and these features are shared by distantly related as well as closely related proteins

Table 4. Performance of Support Vector Machine Prediction of Functional Classes of Novel Proteins

Protein Group and Year of Report	No. of Proteins or Protein Pairs	Percentage of Correctly Predicted Proteins or Protein Pairs	Examples of Correctly Predicted Proteins or Protein Pairs	Examples of Incorrectly Predicted Proteins or Protein Pairs
Enzymes without a homolog in NR databases 2004 [57]	12	66.7%	Thiocyanate hydrolase beta subunit (EC 3.5.5.8) [O66186] Potential cysteine protease avirulence protein avrPpiC2(EC 3.4.22.-) [Q9F3T4] Extracellular phospholipase (EC 3.1.1.5) [P82476]	Extracellular phospholipase (EC 3.1.1.5) [P82476] Alginate lyase precursor (EC4.2.2.3) [P39049]
Enzymes without a homolog in Swissprot database 2004 [57]	50	72%	DNA polymerase III, theta subunit (EC 2.7.7.7) [P28689] Telomere elongation protein (EC2.7.7.-) [P17214] Ammonia monooxygenase (EC 1.13.12.-) [Q04508]	Beta-agarase B (EC 3.2.1.81) [P48840] Alpha-N-AFase II (EC 3.2.1.55) [P82594]
Viral proteins without a homolog in Swissprot database 2004 [124]	25	72%	Endonuclease II [P07059] Outer capsid protein VP4 [P35746] Protein kinase [P00513]	TRL10 (Structural envelop glycoprotein) [AAL27474] BARF0 protein [Q8AZJ4]
Bacterial proteins without a homolog in Swissprot database 2004 [123]	90	76.7%	2-aminomuconate deaminase [P81593] Aminopeptidase G [Q54340]	Alginate lyase [Q59478] Alpha-N-AFase II [P82594]
Plant proteins without a homolog in Swissprot database [252]	31	71.4%	Antimicrobial peptide 4 [AAL05055] Sucrose phosphatase [Q84ZX9]	LeMan3 [Q9FUQ6] MAN5 [Q6YM50]
Pairs of homologous enzymes of different families 2004 [57]	8	62%	Glycolateoxidase [P05414] and IPP isomerase [Q8PW37] Creatine amidinohydrolase [P38488] and Prolinedipeptidase [O58885]	Cystathionine gammasynthase [P38675] and Methionine gamma-lyase [P13254] Exocellobiohydrolase 1 [P38676] and Cystathionine gamma-lyase [Q8VCN5]
Remote homologs [121] from FSSP database [211] 2005	445	46.3%	1cem (1,4-D-glucan-glucanohydrolase catalytic domain) and it's remote homolog 1qazA (Alginate lyase A1-III from *Sphingomonas* Species; Chain: A;)	

of the same functional property [125]. Some of these function-related features might be captured by the residue properties such as hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structures and solvent accessibility [126, 127], which have been incorporated in the descriptors used in the construction of the feature vectors for these proteins.

The function of a protein is determined by a variety of factors. Changes such as local active-site mutation, variations in surface loops, and recruitment of additional domains may result in functional diversity among homologous proteins [23]. While these changes appear to be small at the local sequence level, some of the aspects of these changes may also be captured by the descriptors associated with hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility.

Performance for Predicting Proteins with Specific Structural Characteristics

Subgroups of proteins of specific functional classes are known to have common structural features. For instance, a

number of RNA-binding proteins have a modular structure and contain RNA-binding domains of 70-150 amino acids that mediate RNA recognition [128, 129]. Three classes of RNA-binding domains have been documented to bind RNA in a sequence independent manner, and these domains are RNA-recognition motif (RRM), double-stranded RNA-binding motif (dsRM), and K-homology (KH) domain [129]. A fourth class of RNA-binding domain, S1 RNA-binding domain, has also been found in a number of RNA-associated proteins [130]. These domains have distinguished structural features responsible for RNA recognition and binding. Thus the performance of SVM classification of functional classes of proteins can be evaluated by examining whether or not proteins containing one of these domains can be correctly classified into the respective class [33, 42, 131, 132].

A search of protein family and sequence databases shows that there are a total of 260, 74, 190, and 41 RNA-binding protein sequences known to contain RRM, dsRM, KH and S1 RNA-binding domain respectively. The majority of these sequences are included in the training and testing set of all RNA-binding proteins. In the corresponding independent evaluation set, there are 35, 16, 93, and 10 sequences containing RRM, dsRM, KH, and S1 RNA-binding domain respectively. All but one protein sequence are correctly classified as RNA-binding by SVM, which shows the capability of SVM [33]. The only incorrectly predicted protein sequence is HnRNP-E2 protein fragment in the group that contains KH domain. The incompleteness of this sequence might partially contribute to its incorrect prediction by SVM.

In another example, some lipid-binding proteins are known to contain lipid-binding domains or motifs [79]. Several families of such lipid-binding proteins have been documented, and examples of these families are TIM, PP-binding or GCV_H. These families have distinguished structural features responsible for lipid recognition and binding. A search of protein family and sequence databases shows that there are 227, 184, and 139 lipid-binding protein sequences known to contain TIM, PP-binding or GCV_H domain respectively. The majority of these sequences are included in the training and testing set of all lipid-binding proteins. In the corresponding independent evaluation set, there are 81, 27, and 30 sequences containing TIM, PP-binding or GCV_H domain respectively. Most of these protein sequences are correctly classified as lipid-binding by SVM, and there is only 1, 1, and 2 misclassified sequences in the TIM, PP-binding or GCV_H domain families respectively 42]. The incorrectly predicted protein sequences are triosephosphate isomerase (fragment), putative acyl carrier protein, mitochondrial precursor, glycine cleavage system H protein, mitochondrial precursor (fragment), probable glycine cleavage system H protein 2, mitochondrial precursor. Most of these incorrectly predicted sequences are fragment. Therefore, sequence incompleteness appears to be a factor that partially contributes to the incorrect prediction of these sequences by SVM.

Effect of Different Sets of Protein Descriptors to the Classification of Functional Classes of Proteins

As shown in Table **2** and Table **3**, different sets of protein descriptors have been used in SVM prediction of various functional classes of proteins and peptides, all of which have shown impressive predictive performances [133-135]. Nonthe-less, there is a need to comparatively evaluate the effectiveness of these descriptor-sets in a single study and to examine whether combined use of these descriptor-sets help to improve predictive performance. For such as purpose, we tested the performance of seven popular descriptor-sets and two of their combinations in SVM prediction of six different classes of proteins. These sets are amino acid composition [133] (class 1), dipeptide composition [134] (class 2), normalized Moreau–Broto autocorrelation [136, 137] (class 3), Moran autocorrelation [138] (class 4), Geary autocorrelation [139] (class 5), sets of composition, transition and distribution of physicochemical properties [140-148] (class 6), and sequence order [149-152] (class 7), the frequently used combination of amino acid composition and dipeptide composition [134] (class 8), and combination of the seven individual sets of descriptors (class 9). The six protein functional classes are enzyme EC2.4 [153], G protein-coupled receptors, transporter TC8.A [154], chlorophyll [155], lipid synthesis proteins involved in lipid synthesis, and rRNA-binding proteins. These classes were selected because of their functional diversity and level of difficulty in achieving high prediction performance. The reported SVM prediction performance for these classes tend to be lower than other classes [142], which are ideal for critically evaluating the effectiveness of different descriptor-sets.

The dataset statistics and SVM performance of the nine descriptor-sets are given in Table **5** and the overall performance scores of these descriptor-sets are given in Table **6**. The overall performance scores are composed of 4 categories defined by the values of MCC of a SVM model: "Exceptional", "Good", "Fair" and "Poor" when MCC is in the range of >0.9, 0.8–0.9, 0.6–0.8, and < 0.6 respectively. Overall, there is no single preferred descriptor-set for all cases. Sets 6, 8 and 9 tend to exhibit higher sensitivity, with the exception of chlorophyll proteins, while classes 1 and 7 tend to be among the lowest ranked. The combined classes 8 and 9 generally give the highest MCC values, again with the exception of chlorophyll proteins, while classes 1 and 7 tend to return the lowest MCC values. These findings are consistent with the results from a reported study that suggest that amino acid composition, polarity, solvent accessibility and charge, are more important than other properties, in order of prominence, for SVM classification of specific protein functional classes [156]. Using the entire set of descriptors (Class 9) does not necessarily always gives better performance, which is consistent with the findings that analysis of the contribution of individual descriptors and the selection of the relevant ones are highly useful for improving SVM prediction performance [157-161].

Contribution of Individual Protein Descriptors to the Classification of Functional Classes of Proteins

While different sets of descriptors have been used to describe physicochemical characteristics of individual proteins [2, 33, 34, 162-164], it has been reported that not all descriptors contribute equally to the classification of proteins, some have been found to play relatively more prominent role than others in specific aspects of proteins [162]. It is therefore of interest to examine which descriptors play more prominent role in classification of specific class of proteins such as

Table 5. Dataset Statistics and Prediction Performance of SVM Prediction of Six Protein Functional Classes by Using Different Descriptor Sets

Protein Functional Family	Descriptor Class	Training Set		Testing Set				Independent Evaluation Set							
				P		N		P			N			Q(%)	MCC
		P	N	TP	FN	TN	FP	TP	FN	Sen(%)	TN	FP	Spec(%)		
EC2.4	1	1249	2120	1154	1	9065	12	724	176	80.4	5064	4	99.9	97.0	0.879
	2	1319	2120	1080	5	8806	1	646	154	82.9	5067	1	100.0	97.4	0.884
	3	1105	1756	1295	4	9166	5	768	132	85.3	5066	2	100.0	97.8	0.911
	4	1239	2221	1161	4	8701	5	756	144	84.0	5067	1	100.0	97.6	0.903
	5	1242	2223	1160	2	8690	14	753	147	83.7	5065	3	99.9	97.5	0.900
	6	1214	2077	1145	45	8846	4	741	159	82.3	5067	1	100.0	97.3	0.893
	7	1293	2624	1072	39	8295	8	696	204	77.3	5065	3	99.9	96.5	0.860
	8	1275	2747	1129	0	8177	3	782	118	86.9	5965	3	99.9	98.0	0.921
	9	1358	3887	1015	31	7040	0	796	104	88.4	5067	1	100.0	98.2	0.930
GPCR	1	1590	7458	1847	1	14166	3	501	12	97.7	6776	62	99.1	99.0	0.927
	2	564	711	1728	3	14121	5	498	15	97.1	6800	38	99.4	99.3	0.946
	3	1169	4628	1122	4	10208	1	491	22	95.7	6800	38	99.4	99.2	0.938
	4	1257	4474	1037	1	10363	0	492	21	95.9	6790	48	99.3	99.1	0.930
	5	1290	4724	997	8	10113	0	487	26	94.9	6795	43	99.4	99.1	0.929
	6	757	2060	1536	2	12777	0	494	19	96.3	6813	25	99.6	99.4	0.951
	7	812	2950	1482	1	11887	0	487	26	94.9	6746	92	98.7	98.4	0.885
	8	1590	7458	693	12	7322	57	503	10	98.1	6780	58	99.2	99.1	0.933
	9	834	4361	1461	0	10476	0	493	20	96.1	6819	19	99.7	99.5	0.959
TC8.A	1	98	8014	9	0	13105	0	17	46	27.0	7962	0	100.0	99.4	0.518
	2	94	7962	50	0	14824	0	41	22	65.1	7962	0	100.0	99.7	0.806
	3	94	7962	53	0	14501	0	42	21	66.7	7962	0	100.0	99.7	0.815
	4	94	7962	47	0	11250	0	37	26	58.7	7962	0	100.0	99.7	0.765
	5	94	7962	47	0	11137	0	37	26	58.7	7962	0	100.0	99.7	0.765
	6	94	7962	64	0	15283	0	44	19	69.8	7962	0	100.0	99.8	0.835
	7	94	7962	59	0	15045	0	43	20	68.3	7962	0	100.0	99.8	0.825
	8	114	810	52	0	15114	0	41	22	65.1	7962	0	100.0	99.7	0.806
	9	103	1077	63	0	14847	0	47	16	74.6	16	0	100.0	99.8	0.863
Chlorophyll	1	523	1559	166	0	14297	0	70	12	85.4	6830	16	99.8	99.6	0.83
	2	440	934	248	1	7927	1	73	9	89.0	6841	5	99.9	99.8	0.91
	3	425	603	264	0	15253	0	77	5	93.9	6841	5	99.9	99.9	0.94
	4	415	574	273	1	15282	0	75	7	91.5	6842	4	99.9	99.8	0.93

(Table 5) cond….

Protein Functional Family	Descriptor Class	Training Set		Testing Set				Independent Evaluation Set							
		P	N	P		N		P			N			Q(%)	MCC
				TP	FN	TN	FP	TP	FN	Sen(%)	TN	FP	Spec(%)		
	5	429	615	259	1	15240	1	75	7	91.5	6843	3	100.0	99.9	0.94
	6	482	946	202	5	14910	0	72	10	87.8	6844	2	100.0	99.8	0.92
	7	394	3337	210	85	12517	2	62	20	75.6	6834	12	99.8	99.5	0.79
	8	399	1273	289	1	14582	1	77	5	93.9	6832	14	99.8	99.7	0.89
	9	458	477	231	0	15379	0	76	6	92.7	6842	4	99.9	99.9	0.93
Lipid synthesis	1	849	2026	705	3	8229	7	476	159	75.0	5882	4	99.9	97.5	0.850
	2	927	2037	629	1	8225	0	507	128	79.8	5886	0	100.0	98.0	0.884
	3	898	2968	659	0	7294	0	509	126	80.2	5886	0	100.0	98.1	0.886
	4	968	3227	588	1	7035	0	493	142	77.6	5886	0	100.0	97.8	0.871
	5	970	3280	586	1	6982	0	491	144	77.3	5886	0	100.0	97.8	0.869
	6	874	2112	681	2	8149	1	525	110	82.7	5884	2	100.0	98.3	0.899
	7	863	2415	692	2	7845	2	512	123	80.6	5883	3	100.0	98.1	0.886
	8	815	1613	740	2	8638	11	525	110	80.7	5879	7	99.9	98.2	0.961
	9	800	3492	757	0	6770	0	541	94	85.2	5886	0	100.0	98.6	0.916
rRNA binding	1	548	579	3390	6	9598	22	1821	90	95.3	4662	6	99.9	98.5	0.964
	2	1133	1225	2811	0	8974	0	1827	84	95.6	4668	0	100.0	98.7	0.969
	3	1126	1638	2816	2	8560	1	1811	100	94.8	4668	0	100.0	98.5	0.963
	4	1337	1958	2697	0	8241	0	1783	128	93.3	4668	0	100.0	98.1	0.953
	5	1372	1976	2572	0	8223	0	1784	127	93.4	4668	0	100.0	98.1	0.953
	6	921	1208	2971	52	8991	0	1824	87	95.5	4668	0	100.0	98.7	0.968
	7	878	2743	3040	26	7442	14	1808	103	97.9	4634	34	99.3	97.9	0.951
	8	810	972	3075	3	9182	2	1848	63	96.7	4668	0	100.0	99.0	0.977
	9	1103	3175	2815	26	7024	0	1805	106	94.5	4668	0	100.0	98.4	0.961

lipid-binding proteins. Contribution of individual descriptors to protein classification has been investigated by separately conducting classification using each feature property [162]. By using the same method, one finds that, in order of prominence, the polarity, hydrophobicity, amino acid composition, and solvent accessibility play more prominent role than other feature properties [42]. Polarity and hydrophobicity have been shown to be important for lipid-protein interactions such that lipid binding sites are located in a hydrophobic and low polarity environment [165]. High-affinity lipid binding site is some proteins appear to be located at sequence segments with specific amino acid composition [166], and specific sequence motifs have been used for predicting lipid-binding proteins [167-171]. A study of apolipophorin-III in lipid-free and phospholipid-bound states showed that lipid-binding involves increased solvent accessibility due to gross tertiary structural reorganization [172]. Therefore, the selected descriptors are consistent with these experimental findings.

Analysis of Descriptor Contributions by Using Feature Selection Method

More rigorous feature selection methods [173-175], such as recursive feature elimination (RFE) [176], have been applied to the SVM classification of functional classes of proteins to select those descriptors most relevant to the prediction of proteins of a particular class [176, 177]. The details of the implementation of this method can be found in the

Table 6. MCC-based Performance Scores of SVM Prediction of Different Protein Functional Classes by Using Different Descriptor Classes

Protein Functional Class	Exceptional > 0.9	Good 0.8–0.9	Fair 0.6–0.8	Poor < 0.6
EC2.4	9, 8, 3, 4, 5	6, 2, 1, 7		
GPCR	9, 6, 2, 3, 8, 4, 5, 1	7		
TC8.A		9, 6, 7, 3, 2, 8	4, 5	1
Chlorophyll	3, 5, 4, 9, 6, 2	8, 1	7	
Lipid synthesis	8, 9	6, 7, 3,2, 4, 5, 1		
rRNA binding	8, 2, 6, 1, 3, 9, 5, 4, 7			

Table 7. Protein Descriptors Important for Characterizing DNA-Binding Proteins as Selected by a Feature Selection Method, Recursive Feature Elimination Method

Descriptor Ranking	Descriptor Index	Structural or Physicochemical Property of Descriptor
1	F168	Solvent accessibility Composition Group 1
2	F166	Secondary structure Group 3 3/4th Distribution
3	F147	Secondary structure Composition Group 1
4	F75	Polarity Group 2 1/4th First Distribution
5	F43	Normalized Van der Waals volume Composition Group 2
6	F155	Secondary structure Group 1 2/4th Distribution
7	F91	Polarizability Group 1 1/4th First Distribution
8	F143	Surface tension Group 3 1/4th First Distribution
9	F171	Solvent accessibility Transition Group 1
10	F126	Surface tension Composition Group 1
11	F87	Polarizability Transition Group 1
12	F145	Surface tension Group 3 3/4th Distribution
13	F15	Composition of R
14	F6	Composition of G
15	F177	Solvent accessibility Group 1 3/4th Distribution
16	F154	Secondary structure Group 1 1/4th First Distribution
17	F89	Polarizability Transition Group 3
18	F133	Surface tension Group 1 1/4th First Distribution
19	F42	Normalized Van der Waals volume Composition Group 1
20	F85	Polarizability Composition Group 2
21	F175	Solvent accessibility Group 1 1/4th First Distribution
22	F130	Surface tension Transition Group 2
23	F127	Surface tension Composition Group 2
24	F151	Secondary structure Transition Group 2
25	F98	Polarizability Group 2 3/4th Distribution

(Table 7) contd….

Descriptor Ranking	Descriptor Index	Structural or Physicochemical Property of Descriptor
26	F8	Composition of I
27	F67	Polarity Transition Group 2
28	F148	Secondary structure Composition Group 2

literature [94, 178]. Feature selection procedure can be demonstrated by the following illustrative example of the development of a SVM classification system for predicting DNA-binding proteins: This system is trained by using a Gaussian kernel function with an adjustable parameter σ. Sequential variation of σ is conducted against the whole training set to find a value that gives the best prediction accuracy. This prediction accuracy is evaluated by means of 5-fold cross-validation. In the first step, for a fixed σ, the SVM classifier is trained by using the complete set of features (protein descriptors) described in the previous section. The second step involves the computation of the ranking criterion score DJ(i) for each feature in the current set. All of the computed DJ(i) is subsequently ranked in descending order. The third step involves the removal the m features with smallest criterion scores. In the fourth step, the SVM classification system is re-trained by using the remaining set of features, and the corresponding prediction accuracy is computed by means of 5-fold cross-validation. The first to fourth steps are then repeated for other values of σ. After the completion of these procedures, the set of features and parameter σ that give the best prediction accuracy are selected.

A total of 28 features were selected by RFE, which are given in Table **7**. In order of prominence, compositions of specific amino acids, Van der Waals volume, polarity, polarizability, surface tension, secondary structure, and solvent accessibility are found to be important for predicting DNA-binding proteins. Protein-DNA binding is known to involve specific recognition sequence and induced conformation changes [179]. Therefore it is expected that the combined features of amino acid composition and surface tension is important for characterizing DNA-binding proteins. DNA binding also involves spatial arrangement or pre-arrangement of specific group of amino acids at the binding site [180]. It is thus not surprising that such important interactions as polarizability, hydrophobicity, polarity and surface tension are coupled to the size of the amino acid sequence segment at a DNA-binding site. Many proteins bind DNA via minor groove interaction between protein non-polar surfaces and DNA hydrophobic sugar clusters [181]. As a result, the combined features of hydrophobicity and solvent accessibility are expected to be important for describing these proteins.

The usefulness of these 28 selected features can be further tested by constructing a SVM classification system based solely on these features. The prediction accuracies of this new system are 87.2% and 92.6% for DNA-binding and non-DNA-binding proteins respectively, which is slightly improved against those of 85.7% and 91.2% by using all features. This suggests that the use of selected subset of features enhances prediction performance by reducing the noise created by the redundant and irrelevant features.

Comparison of SVM Prediction Performance Under Different Kernel Functions

Apart from the Gaussian kernel function of sequence-derived physicochemical properties, several other kernel functions have been developed and applied for SVM classification of proteins and DNAs [182-190]. It is of interest to test the usefulness of some of these kernel functions for predicting functional classes of proteins. The string-kernel function has been extensively used and it has shown promising potential for protein and DNA studies [182, 183]. This kernel function is constructed by comparison of sequences of classes of proteins or DNAs and the assignment of individual weights to amino acids or nucleotides to describe physicochemical or other characteristics of the proteins and DNAs. This kernel function is used to develop three SVM systems for predicting the class of lipid-degradation, lipid metabolism, and lipid synthesis proteins. Spectrum kernel with mismatches 189] is used to generate the string-kernel for each protein. Testing results by using an independent set of proteins for each class show that the SE is 77.2%, 75.8%, 77.8%, and the SP is 97.6%, 96.4%, 94.2% for each of these classes respectively 42]. Thus comparable prediction performance can be achieved by using string-kernel SVM, which suggests the usefulness of this and other kernel functions for SVM prediction of functional classes of proteins.

Comparison of SVM Prediction Performance with Other Machine Learning Methods

Apart from SVM, several other machine learning (ML) methods have been explored for predicting the functional classes of proteins and peptides. These methods include artificial neural network (ANN), k-nearest neighbors (KNN), decision tree and hidden Markov model (HMM). They have been used for predicting enzymes [29], receptors [35], transporters [35], structural proteins [35], mitochondrial proteins [191], cell cycle regulated proteins [192], growth factors [35], and allergen proteins [193, 194]. The reported P+ and P- values of these ML methods are in the range of 37.8%~87% and 66.0%~99.9%, with the majority concentrated in the range of 60%~85% and 70%~90% respectively. These values are slightly lower than the values of 75%~95% and 80%~99.9% of the SVM, suggesting that other ML methods are also useful for predicting the functional class of proteins and peptides.

UNDERLYING DIFFICULTIES IN USING SUPPORT VECTOR MACHINES

The performance of SVM critically depends on the diversity of samples (proteins and peptides) in a training dataset and the appropriate representation of these samples. The datasets used in many of the reported studies are not expected to be fully representative of all of the proteins and peptides with and without a particular functional and interaction profile. Various degrees of inadequate sampling representation likely affect, to a certain extent, the prediction accuracy of the developed machine learning models. SVM is not applicable for proteins, peptides and small molecules with insufficient knowledge about their specific functional and interaction profile. Searching of the information about proteins and peptides known to possess a particular profile and those do not possess that profile is a key to more extensive exploration of statistical learning methods for facilitating the study of protein functional and interaction profiles. Apart from literature sources such as PubMed (http://www.ncbi.nlm.nih.gov/ entrez/query.fcgi?db=PubMed), databases such as Swiss-Prot [195], Genbank [196], pirpsd [197], geneontology [198], PDB [199], enzyme database [200], TransportDB [201], HMTD [202], ABCdb [203], TiPS [204], GPCRDB [205], SYFPEITHI [206], MHCPEP [207], JenPep [208], MHCBN [209], FIMM [210], and FSSP database [211] are also useful for obtaining information about protein/peptide functional and interaction profiles.

In the datasets of some of the reported studies, there appears to be an imbalance between the number of samples having a profile and those without the profile. SVM method tends to produce feature vectors that push the hyper-plane towards the side with smaller number of data [212], which often lead to a reduced prediction accuracy for the class with a smaller number of samples or less diversity than those of the other class. It is however inappropriate to simply reduce the size of non-members to artificially match that of members, since this compromises the diversity needed to fully represent all non-members. Computational methods for readjusting biased shift of hyperplane are being explored [213]. Application of these methods may help improving the prediction accuracy of SVM in the cases involving imbalanced data.

While a number of descriptors have been introduced for representing proteins and peptides [2, 31, 34, 52], most reported studies typically use only a portion of these descriptors. It has been found that, in some cases, selection of a proper subset of descriptors is useful for improving the performance of SVM [173-175]. Therefore, there is a need to explore different combination of descriptors and to select more optimum set of descriptors for more cases, which can be conducted by using feature selection methods [173-175]. Effort also been directed at the improvement of the efficiency and speed of feature selection methods [214], which will enable a more extensive application of feature selection methods. Moreover, indiscriminate use of the existing descriptors, particularly those of overlapping and redundant descriptors, may introduce noise as well as extending the coverage of some the aspects of these special features. Thus, it may be necessary to introduce new descriptors for the systems that have been described by overlapping and redundant descriptors. Investigation of cases of incorrectly predicted samples have also suggested that the currently-used descriptors may not always be sufficient for fully representing the structural and physicochemical properties of proteins, peptides and small molecules [215-217]. These have prompted works for developing new descriptors [41].

CONCLUDING REMARKS

SVM has consistently shown promising capability for predicting functional classes of proteins and peptides. Proper use of descriptors for representing proteins and peptides may help further improving the performance of SVM for predicting functional profiles of proteins and peptides. The introduction of new descriptors would better represent characteristics that correlate with novel functional and interaction profiles. Moreover, various feature selection methods may be used for selecting optimal set of descriptors for a particular prediction problem. Existing algorithms can be improved and new algorithms may be introduced for enhancing the performance and accuracy of SVM. The prediction capability of SVM can be further enhanced with increasing availability of biological data and more extensive knowledge about sequence, structure, transcription, post-transcriptional processing features that define the functional profiles of proteins and peptides. These efforts will enable the development of SVM into useful tools for facilitating the study of functional profiles of proteins and peptides to complement other well-established methods such as sequence similarity and clustering methods.

REFERENCES

[1] Lo, S.L., Cai, C.Z., Chen, Y.Z. and Chung, M.C. (**2005**) *Proteomics*, *5*, 876-884.
[2] Bock, J.R. and Gough, D.A. (**2001**) *Bioinformatics*, *17*, 455-460.
[3] Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. (**2000**) *Nature*, *405*, 823-826.
[4] Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (**1998**) *J. Mol. Biol.*, *283*, 707-725.
[5] Godzik, A., Jambon, M. and Friedberg, I. (**2007**) *Cell Mol. Life Sci.*, *64*, 2505-2511.
[6] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. and Singh, M. (**2005**) *Bioinformatics*, *21 Suppl 1*, i302-310.
[7] Sharan, R., Ulitsky, I. and Shamir, R. (**2007**) *Mol. Syst. Biol.*, *3*, 88.
[8] Schuler, G.D. (**1998**) *Methods Biochem. Anal.*, *39*, 145-171.
[9] Bork, P. and Koonin, E.V. (**1998**) *Nat. Genet.*, *18*, 313-318.
[10] Baxevanis, A.D. (**1998**) *Methods Biochem. Anal.*, *39*, 172-188.
[11] Hodges HC, T.J. (**2002**) *FASB J*, *16*:A543-A543.
[12] Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (**2002**) *Nucleic Acids Res.*, *30*, 1575-1584.
[13] Enright, A.J. and Ouzounis, C.A. (**2000**) *Bioinformatics*, *16*, 451-457.
[14] Fujiwara, Y. and Agsogwa, M. (**2002**) *NEC Res. Dev.*, *43*, 238-241.
[15] Eisen, J.A. (**1998**) *Genome Res.*, *8*, 163-167.
[16] Benner, S.A., Chamberlin, S.G., Liberles, D.A., Govindarajan, S. and Knecht, L. (**2000**) *Res. Microbiol.*, *151*, 97-106.
[17] Whisstock, J.C. and Lesk, A.M. (**2003**) *Q. Rev. Biophys.*, *36*, 307-340.
[18] Rost, B. (**2002**) *J. Mol. Biol.*, *318*, 595-608.
[19] Shah, I. and Hunter, L. (**1997**) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, *5*, 276-283.
[20] Han, L.Y., Cai, C.Z., Ji, Z.L., Cao, Z.W., Cui, J. and Chen, Y.Z. (**2004**) *Nucleic Acids Res.*, *32*, 6437-6444.
[21] Smith, T.F. and Zhang, X. (**1997**) *Nat. Biotechnol.*, *15*, 1222-1223.
[22] Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y. and Chen, Y. (**2006**) *Proteomics*, *6*, 4023-4037.
[23] Todd, A.E., Orengo, C.A. and Thornton, J.M. (**2001**) *J. Mol. Biol.*, *307*, 1113-1143.

[24] Teichmann, S.A., Murzin, A.G. and Chothia, C. (**2001**) *Curr. Opin. Struct. Biol., 11*, 354-363.
[25] Aravind, L. (**2000**) *Genome Res., 10*, 1074-1077.
[26] Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (**1999**) *Nature, 402*, 86-90.
[27] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (**1999**) *Science, 285*, 751-753.
[28] Cai, Y.D. and Lin, S.L. (**2003**) *Biochim. Biophys. Acta., 1648*, 127-133.
[29] Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H.H., Rapacki, K., Workman, C., Andersen, C.A., Knudsen, S., Krogh, A., Valencia, A. and Brunak, S. (**2002**) *J. Mol. Biol., 319*, 1257-1265.
[30] Cai, C.Z., Han, L.Y., Ji, Z.L. and Chen, Y.Z. (**2004**) *Proteins, 55*, 66-76.
[31] Karchin, R., Karplus, K. and Haussler, D. (**2002**) *Bioinformatics, 18*, 147-159.
[32] des Jardins, M., Karp, P.D., Krummenacker, M., Lee, T.J. and Ouzounis, C.A. (**1997**) *Proc. Int. Conf. Intell. Syst. Mol. Biol., 5*, 92-99.
[33] Han, L.Y., Cai, C.Z., Lo, S.L., Chung, M.C. and Chen, Y.Z. (**2004**) *RNA, 10*, 355-368.
[34] Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. and Chen, Y.Z. (**2003**) *Nucleic Acids Res., 31*, 3692-3697.
[35] Jensen, L.J., Gupta, R., Staerfeldt, H.H. and Brunak, S. (**2003**) *Bioinformatics, 19*, 635-642.
[36] Bhasin, M. and Raghava, G.P. (**2004**) *J. Biol. Chem., 279*, 23262-23266.
[37] Cai, Y.D. and Chou, K.C. (**2005**) *J. Proteome Res., 4*, 967-971.
[38] Guo, Y.Z., Li, M., Lu, M., Wen, Z., Wang, K., Li, G. and Wu, J. (**2006**) *Amino Acids, 30*, 397-402.
[39] Bhasin, M. and Raghava, G.P. (**2004**) *Nucleic Acids Res., 32*, W383-389.
[40] Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. and Suwa, M. (**2005**) *Nucleic Acids Res., 33*, W148-153.
[41] Bhardwaj, N., Langlois, R.E., Zhao, G. and Lu, H. (**2005**) *Nucleic Acids Res., 33*, 6486-6493.
[42] Lin, H.H., Han, L.Y., Zhang, H.L., Zheng, C.J., Xie, B. and Chen, Y.Z. (**2006**) *J. Lipid Res., 47*, 824-831.
[43] Dobson, P.D. and Doig, A.J. (**2005**) *J. Mol. Biol., 345*, 187-199.
[44] Han, L.Y., Zheng, C.J., Xie, B., Jia, J., Ma, X.H., Zhu, F., Lin, H.H., Chen, X. and Chen, Y.Z. (**2007**) *Drug Discov. Today, 12*, 304-313.
[45] Huang, N., Chen, H. and Sun, Z. (**2005**) *Protein Eng. Des. Sel., 18*, 365-368.
[46] Xu, J.R., Zhang, J.X., Han, B.C., Liang, L. and Ji, Z.L. (**2007**) *Nucleic Acids Res., 35*, W538-542.
[47] Bhasin, M. and Raghava, G.P. (**2004**) *Vaccine, 22*, 3195-3204.
[48] Smialowski, P., Schmidt, T., Cox, J., Kirschner, A. and Frishman, D. (**2006**) *Proteins, 62*, 343-355.
[49] Kumar, M., Verma, R. and Raghava, G.P. (**2005**) *J. Biol. Chem.*
[50] Gromiha, M.M. and Suresh, M.X. (**2007**) *Proteins.*
[51] Cai, Y.D. and Doig, A.J. (**2004**) *Bioinformatics, 20*, 1292-1300.
[52] Reczko, M. and Bohr, H. (**1994**) *Nucleic Acids Res., 22*, 3616-3619.
[53] Hediger, M.A. (**1994**) *J. Exp. Biol., 196*, 15-49.
[54] Borst, P. and Elferink, R.O. (**2002**) *Annu. Rev. Biochem., 71*, 537-592.
[55] Seal, R.P. and Amara, S.G. (**1999**) *Annu. Rev. Pharmacol. Toxicol., 39*, 431-456.
[56] Joet, T., Morin, C., Fischbarg, J., Louw, A.I., Eckstein-Ludwig, U., Woodrow, C. and Krishna, S. (**2003**) *Expert. Opin. Ther. Targets, 7*, 593-602.
[57] Birch, P.J., Dekker, L.V., James, I.F., Southan, A. and Cronk, D. (**2004**) *Drug Discov. Today, 9*, 410-418.
[58] Dutta, A.K., Zhang, S., Kolhatkar, R. and Reith, M.E. (**2003**) *Eur. J. Pharmacol., 479*, 93-106.
[59] Lee, W. and Kim, R.B. (**2004**) *Annu. Rev. Pharmacol. Toxicol., 44*, 137-166.
[60] Kunta, J.R. and Sinko, P.J. (**2004**) *Curr. Drug Metab., 5*, 109-124.
[61] Driessen, A.J., Rosen, B.P. and Konings, W.N. (**2000**) *Trends Biochem. Sci., 25*, 397-401.
[62] Saier, M.H., Jr. (**2000**) *Microbiol. Mol. Biol. Rev., 64*, 354-411.
[63] Lewin, B. (**2000**) *Genes VII* (Oxford: Oxford University Press).
[64] Sarai, A. and Kono, H. (**2005**) *Annu. Rev. Biophys. Biomol. Struct., 34*, 379-398.
[65] Aguilar, D., Oliva, B., Aviles, F.X. and Querol, E. (**2002**) *Bioinformatics, 18*, 597-607.
[66] Stawiski, E.W., Gregoret, L.M. and Mandel-Gutfreund, Y. (**2003**) *J. Mol. Biol., 326*, 1065-1079.
[67] Garvie, C.W. and Wolberger, C. (**2001**) *Mol. Cell, 8*, 937-946.
[68] Bewley, C.A., Gronenborn, A.M. and Clore, G.M. (**1998**) *Annu. Rev. Biophys. Biomol. Struct., 27*, 105-131.
[69] Luscombe, N.M. and Thornton, J.M. (**2002**) *J. Mol. Biol., 320*, 991-1009.
[70] Luscombe, N.M., Laskowski, R.A. and Thornton, J.M. (**2001**) *Nucleic Acids Res., 29*, 2860-2874.
[71] Fujii, Y., Shimizu, T., Toda, T., Yanagida, M. and Hakoshima, T. (**2000**) *Nat. Struct. Biol., 7*, 889-893.
[72] Steffen, N.R., Murphy, S.D., Tolleri, L., Hatfield, G.W. and Lathrop, R.H. (**2002**) *Bioinformatics, 18 Suppl 1*, S22-30.
[73] Downes, C.P., Gray, A. and Lucocq, J.M. (**2005**) *Trends Cell Biol., 15*, 259-268.
[74] Glatz, J.F., Luiken, J.J., van Bilsen, M. and van der Vusse, G.J. (**2002**) *Mol. Cell Biochem., 239*, 3-7.
[75] Haunerland, N.H. and Spener, F. (**2004**) *Prog. Lipid Res., 43*, 328-349.
[76] Bingle, C.D. and Craven, C.J. (**2004**) *Trends Immunol., 25*, 53-55.
[77] Bernlohr, D.A., Simpson, M.A., Hertzel, A.V. and Banaszak, L.J. (**1997**) *Annu. Rev. Nutr., 17*, 277-303.
[78] Niggli, V. (**2001**) *Trends Biochem. Sci., 26*, 604-611.
[79] Balla, T. (**2005**) *J. Cell Sci., 118*, 2093-2104.
[80] Pebay-Peyroula, E. and Rosenbusch, J.P. (**2001**) *Curr. Opin. Struct. Biol., 11*, 427-432.
[81] Fyfe, P.K., Hughes, A.V., Heathcote, P. and Jones, M.R. (**2005**) *Trends Plant Sci., 10*, 275-282.
[82] Bolanos-Garcia, V.M. and Miguel, R.N. (**2003**) *Prog. Biophys. Mol. Biol., 83*, 47-68.
[83] Hanhoff, T., Lucke, C. and Spener, F. (**2002**) *Mol. Cell Biochem., 239*, 45-54.
[84] Weisiger, R.A. (**2002**) *Mol. Cell Biochem., 239*, 35-43.
[85] Palsdottir, H. and Hunte, C. (**2004**) *Biochim. Biophys. Acta, 1666*, 2-18.
[86] Lichtman, A.K.A.A.H. (**2005**) *Cellular and Molecular Immunology*, Updated Edition (Book + Student Consult + Evolve, Volume -, 5th edition Edition (-: W.B. Saunders Company).
[87] Shoshan, S.H. and Admon, A. (**2004**) *Pharmacogenomics, 5*, 845-859.
[88] Matsumura, M., Fremont, D.H., Peterson, P.A. and Wilson, I.A. (**1992**) *Science, 257*, 927-934.
[89] Zhang, C., Anderson, A. and DeLisi, C. (**1998**) *J. Mol. Biol., 281*, 929-947.
[90] McFarland, B.J. and Beeson, C. (**2002**) *Med. Res. Rev., 22*, 168-203.
[91] Bock, J.R. and Gough, D.A. (**2003**) *Bioinformatics, 19*, 125-134.
[92] Martin, S., Roe, D. and Faulon, J.L. (**2005**) *Bioinformatics, 21*, 218-226.
[93] Ben-Hur, A. and Noble, W.S. (**2005**) *Bioinformatics, 21 Suppl 1*, i38-i46.
[94] Xue, Y., Yap, C.W., Sun, L.Z., Cao, Z.W., Wang, J.F. and Chen, Y.Z. (**2004**) *J. Chem. Inf. Comput. Sci., 44*, 1497-1505.
[95] Li, Z.R., Lin, H.H., Han, L.Y., Jiang, L., Chen, X. and Chen, Y.Z. (**2006**) *Nucleic Acids Res., 34*, W32-37.
[96] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. (**2005**) *Protein Identification and Analysis Tools on the ExPASy Serve*r. In *The Proteomics Protocols Handbook*, John, M.W. ed. (Humana Press), pp. 571-607.
[97] Kawashima, S. and Kanehisa, M. (**2000**) *Nucleic Acids Res., 28*, 374.
[98] Cid, H., Bunster, M., Canales, M. and Gazitua, F. (**1992**) *Protein Eng., 5*, 373-375.
[99] Bhaskaran, R. and Ponnuswammy, P.K. (**1988**) *Int. J. Pept. and Protein Res., 32*, 242-255.
[100] Charton, M. and Charton, B.I. (**1982**) *J. Theor. Biol., 99*, 629-644.
[101] Chothia, C. (**1976**) *J. Mol. Biol., 105*, 1-12.
[102] Bigelow, C.C. (**1967**) *J. Theor. Biol., 16*, 187-211.
[103] Charton, M. (**1981**) *J. Theor. Biol., 91*, 115-123.
[104] Dayhoff, H. and Calderone, H. (**1978**) *Altas of Protein Sequence and Structure, 5*, 363-373.
[105] Schneider, G. and Wrede, P. (**1994**) *Biophys. J., 66*, 335-344.
[106] Chou, K.C. (**2000**) *Biochem. Biophys. Res. Commun., 278*, 477-483.

[107] Chou, K.C. and Cai, Y.D. (**2004**) *Biochem. Biophys. Res. Commun., 320*, 1236-1239.
[108] Grantham, R. (**1974**) *Science, 185*, 862-864.
[109] Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. and Kim, S.H. (**1999**) *Proteins, 35*, 401-407.
[110] Dubchak, I., Muchnik, I., Holbrook, S.R. and Kim, S.H. (**1995**) *Proc. Natl. Acad. Sci. USA, 92*, 8700-8704.
[111] Vapnik, V.N. (**1995**) *The nature of statistical learning theory* (New York: Springer).
[112] Lei, Z. and Dai, Y. (**2006**) *BMC Bioinformatics, 7*, 491.
[113] Plewczynski, D., Tkacz, A., Godzik, A. and Rychlewski, L. (**2005**) *Cell. Mol. Biol. Lett., 10*, 73-89.
[114] Weiss, S.M. and Kulikowski, C.A. (**1991**) Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems (San Francisco, CA, USA, Morgan Kaufmann Publishers Inc).
[115] Shao, J. and Tu, D. (**1995**) *The Jackknife and Bootstrap* (New York, NY, USA: Springer).
[116] Liu, H., Yang, J., Wang, M., Xue, L. and Chou, K.C. (**2005**) *Protein J., 24*, 385-389.
[117] Wang, M.L., Yao, H. and Xu, W.B. (**2005**) *Comput. Biol. Chem., 29*, 95-100.
[118] Johnson, S.C. (**1967**) *Psychometrika, 32*, 241-254.
[119] Provost, F., Fawcett, T. and Kohavi, R. (**1998**) The case against accuracy estimation for comparing induction algorithms. In Proc. 15th International Conf. on Machine Learning. (San Francisco, CA: Morgan Kaufmann), pp. 445-453.
[120] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., and Nielsen, H. (**2000**) *Bioinformatics, 16*, 412-424.
[121] Zhang, Z., Kochhar, S. and Grigorov, M.G. (**2005**) *Protein Sci., 14*, 431-444.
[122] Hou, Y., Hsu, W., Lee, M.L. and Bystroff, C. (**2004**) *Proteins., 57*, 518-530.
[123] Cui, J., Han, L.Y., Cai, C.Z., Zheng, C.J., Ji, Z.L. and Chen, Y.Z. (**2005**) *J. Mol. Microbiol. Biotechnol., 9*, 86-100.
[124] Han, L.Y., Cai, C.Z., Ji, Z.L. and Chen, Y.Z. (**2005**) *Virology, 331*, 136-143.
[125] Schomburg, I., Chang, A. and Schomburg, D. (**2002**) *Nucleic Acids Res., 30*, 47-49.
[126] Bull, H.B. and Breese, K. (**1974**) *Arch. Biochem. Biophys., 161*, 665-670.
[127] Lin, T.Y. and Timasheff, S.N. (**1996**) *Protein Sci., 5*, 372-381.
[128] Mattaj, I.W. (**1993**) *Cell, 73*, 837-840.
[129] Perez-Canadillas, J.M. and Varani, G. (**2001**) *Curr. Opin. Struct. Biol., 11*, 53-58.
[130] Bycroft, M., Hubbard, T.J., Proctor, M., Freund, S.M. and Murzin, A.G. (**1997**) *Cell, 88*, 235-242.
[131] Kunik, V., Solan, Z., Edelman, S., Ruppin, E. and Horn, D. (**2005**) *Proc. IEEE Comput. Syst. Bioinform. Conf.*, 80-85.
[132] Leslie, C.S., Eskin, E., Cohen, A., Weston, J. and Noble, W.S. (**2004**) *Bioinformatics, 20*, 467-476.
[133] Chou, K. and Cai, Y. (**2005**) *J. Chem. Inf. Model., 45*, 407-413.
[134] Gao, Q., Wang, Z., Yan, C. and Du, Y. (**2005**) *FEBS Lett., 20*, 16.
[135] Li, Z., Lin, H., Han, L., Jiang, L., Chen, X. and Chen, Y. (**2006**) *Nuclei Acid Res., 34*, W32-37.
[136] Feng, Z. and Zhang, C. (**2000**) *J. Protein Chem., 19*, 262-275.
[137] Lin, Z. and Pan, X. (**2001**) *J. Protein Chem., 20*, 217-220.
[138] Horne, D. (**1988**) *Biopolymers, 27*, 451-477.
[139] Sokal, R. and Thomson, B. (**2006**) *Am. J. Phys. Anthropol., 129*, 121-131.
[140] Bock, J. and Gough, D. (**2001**) *Bioinformatics, 17*, 455-460.
[141] Cai, C., Han, L., Ji, Z., Chen, X. and Chen, Y. (**2003**) *Nuclei Acid Res., 31*, 3692-3697.
[142] Cai, C., Han, L., Ji, Z. and Chen, Y. (**2004**) *Proteins, 55*, 66-76.
[143] Cui, J., Han, L., Lin, H., Zhang, H., Tang, Z., Zheng, C., Cao, Z. and Chen, Y. (**2006**) *Mol. Immunol., 43*.
[144] Dubchak, I., Muchnik, M., Holbrook, S., and Kim, S. (**1995**) *Proc. Natl. Acad. Sci. USA, 92*, 8700-8704.
[145] Dubchak, I., Muchnick I, Mayor, C., Dralyuk, I. and Kim, S. (**1999**) *Proteins, 35*, 401-407.
[146] Han, L., Cai, C., Lo, S., Chung, M. and Chen, Y. (**2004**) *RNA, 10*, 355-368.
[147] Lo, S., Cai, C., Chen, Y. and Chung, M. (**2005**) *Proteomics, 5*, 876-884.
[148] Lin, H., Han, L., Cai, C., Ji, Z. and Chen, Y. (**2006**) *Proteins, 62*, 218-231.
[149] Chou, K. (**2000**) *Biochem. Biophys. Res. Commun., 278*, 477-483.
[150] Chou, K. and Cai, Y. (**2004**) *Biochem. Biophys. Res. Commun., 320*, 1236-1239.
[151] Grantham, R. (**1974**) *Science, 185*, 862-864.
[152] Schneider, G. and Wrede, P. (**1994**) *Biophys. J., 66*, 355-344.
[153] NC-IUBMB (**1992**) *Enzyme Nomenclature* (San Diego, California: Academic Press).
[154] Saier, M.H.J., Tran, C.V. and Barabote, R.D. (**2006**) *Nuclei Acid Res., 34*, D181-D186.
[155] Suzuki, J., Bollivar, D. and Bauer, C. (**1997**) *Annu. Rev. Genet., 31*, 61-89.
[156] Lin, H., Han, L., Zhang, H., Zheng, C., Xie, B. and Chen, Y. (**2006**).
[157] Glen, W., Dunn, W. and Scott, R. (**1989**) *Tetrahedron Comput. Methodol., 2*, 349-376.
[158] Xue, L. and Bajorath, J. (**2000**) *Comb. Chem.High Throughput Screen, 3*, 363-372.
[159] Xue, L., Godden, J. and Bajorath, J. (**1999**) *J. Chem. Inf. Comput. Sci., 39*, 669-704.
[160] Xue, L., Godden, J. and Bajorath, J. (**2000**) *Comb. Chem. High Throughput Screen, 3*, 363-372.
[161] Xue, L., Godden, J. and Bajorath, J. (**2000**) *J. Chem. Inf. Comput. Sci., 40*, 1227-1234.
[162] Ding, C.H. and Dubchak, I. (**2001**) *Bioinformatics, 17*, 349-358.
[163] Cai, Y.D., Liu, X.J., Xu, X.B. and Chou, K.C. (**2002**) *J. Comput. Chem., 23*, 267-274.
[164] Cai, Y.D., Liu, X.J., Xu, X.B. and Chou, K.C. (**2002**) *Comput. Chem., 26*, 293-296.
[165] Lugo, M.R. and Sharom, F.J. (**2005**) *Biochemistry, 44*, 643-655.
[166] Hamilton, S.E., Recny, M. and Hager, L.P. (**1986**) *Biochemistry, 25*, 8178-8183.
[167] Gonnet, P. and Lisacek, F. (**2002**) *Bioinformatics, 18*, 1091-1101.
[168] Eisenhaber, F., Eisenhaber, B., Kubina, W., Maurer-Stroh, S., Neuberger, G., Schneider, G. and Wildpaner, M. (**2003**) *Nucleic Acids Res., 31*, 3631-3634.
[169] Juncker, A.S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. and Krogh, A. (**2003**) *Protein Sci., 12*, 1652-1662.
[170] Gonnet, P., Rudd, K.E. and Lisacek, F. (**2004**) *Proteomics, 4*, 1597-1613.
[171] Eisenhaber, B., Eisenhaber, F., Maurer-Stroh, S. and Neuberger, G. (**2004**) *Proteomics, 4*, 1614-1625.
[172] Raussens, V., Narayanaswami, V., Goormaghtigh, E., Ryan, R.O. and Ruysschaert, J.M. (**1996**) *J. Biol. Chem., 271*, 23089-23095.
[173] Xue, Y., Li, Z.R., Yap, C.W., Sun, L.Z., Chen, X. and Chen, Y.Z. (**2004**) *J. Chem. Inf. Comput. Sci., 44*, 1630-1638.
[174] Al-Shahib, A., Breitling, R. and Gilbert, D. (**2005**) *Int. J. Neural. Syst., 15*, 259-275.
[175] Al-Shahib, A., Breitling, R. and Gilbert, D. (**2005**) *Appl. Bioinformatics, 4*, 195-203.
[176] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (**2002**) *Machine Learning, 46*, 389-422.
[177] Yu, H., Yang, J., Wang, W. and Han, J. (**2003**) In *IEEE Computer Society Bioinformatics Conference (CSB'03)* 220-228: Stanford, California.
[178] Xue, Y., Li, Z.R., Yap, C.W., Sun, L.Z., Chen, X. and Chen, Y.Z. (**2004**) *J. Chem. Infor. Comp. Sci., 44*, 1630-1638.
[179] Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W. and Roberts, R.J. (**1993**) *Cell, 74*, 299-307.
[180] Patel, A., Shuman, S. and Mondragon, A. (**2006**) *J. Biol. Chem., 281*, 6030-6037.
[181] Tolstorukov, M.Y., Jernigan, R.L. and Zhurkin, V.B. (**2004**) *J. Mol. Biol., 337*, 65-76.
[182] Vishwanathan, S.V.N. and Smola, A.J. (**2002**) In Proceedings of Neural Information Processing Systems, 2002.
[183] Ratsch, G., Sonnenburg, S. and Scholkopf, B. (**2005**) *Bioinformatics, 21 Suppl 1*, i369-i377.
[184] Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, K.R. (**2000**) *Bioinformatics, 16*, 799-807.
[185] Jaakkola, T., Diekhans, M. and Haussler, D. (**1999**) Using the Fisher Kernel Method to Detect Remote Protein Homologies. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes and R. Zirnmer, eds.) pp. 149-158, Menlo Park, CA: AAAI Press.
[186] Tsuda, K., Kawanabe, M., Ratsch, G., Sonnenburg, S. and Muller, K.R. (**2002**) *Neural. Comput., 14*, 2397-2414.

[187] Liao, L. and Noble, W.S. (**2003**) *J. Comput. Biol.*, *10*, 857-868.
[188] Vert, J.-P., Saigo, H. and Akutsu, T. (**2003**) *Local alignment kernels for biological sequences.* In *Kernel Methods in Computational Biology*, pp. 131-154, Cambridge: MIT Press.
[189] Leslie, C., Kuang, R. and Eskin, E. (**2003**) Inexact matching string kernels for protein classification. In Kernel Methods in Computational Biology, pp. 95-112, Cambridge: MIT Press.
[190] Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y. and Leslie, C. (**2005**) *J. Bioinform. Comput. Biol.*, *3*, 527-550.
[191] Kumar, M., Verma, R. and Raghava, G.P. (**2006**) *J. Biol. Chem.*, *281*, 5357-5363.
[192] de Lichtenberg, U., Jensen, T.S., Jensen, L.J. and Brunak, S. (**2003**) *J. Mol. Biol.*, *329*, 663-674.
[193] Zorzet, A., Gustafsson, M. and Hammerling, U. (**2002**) In *Silico. Biol.*, *2*, 525-534.
[194] Soeria-Atmadja, D., Zorzet, A., Gustafsson, M.G. and Hammerling, U. (**2004**) *Int. Arch. Allergy Immunol.*, *133*, 101-112.
[195] Dorazilova, V. and Vedralova, J. (**1992**) *Cesk Patol.*, *28*, 245-247.
[196] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (**2004**) *Nucleic Acids Res.*, *32*, D23-26.
[197] Barker, W.C., Garavelli, J.S., McGarvey, P.B., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.S., Ledley, R.S., Mewes, H.W., Pfeiffer, F., Tsugita, A. and Wu, C. (**1999**) *Nucleic Acids Res.*, *27*, 39-43.
[198] Chalmel, F., Lardenois, A., Thompson, J.D., Muller, J., Sahel, J.A., Leveillard, T. and Poch, O. (**2005**) *Bioinformatics*, *21*, 2095-2096.
[199] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (**2000**) *Nucleic Acids Res.*, *28*, 235-242.
[200] Bairoch, A. (**2000**) *Nucleic Acids Res.*, *28*, 304-305.
[201] Ren, Q., Kang, K.H. and Paulsen, I.T. (**2004**) *Nucleic Acids Res.*, *32*, D284-288.
[202] Yan, Q. and Sadee, W. (**2000**) *AAPS Pharm. Sci.*, *2*, E20.
[203] Quentin, Y. and Fichant, G. (**2000**) *J. Mol. Microbiol. Biotechnol.*, *2*, 501-504.
[204] Alexander, S., Peters, J., Mead, A. and Lewis, S. (**1999**) *Trends Pharmacol. Sci.*, *19*, 5-85.
[205] Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F.E. and Vriend, G. (**2003**) *Nucleic Acids Res.*, *31*, 294-297.
[206] Rammensee, H., Bachmann, J., Emmerich, N.P., Bachor, O.A. and Stevanovic, S. (**1999**) *Immunogenetics*, *50*, 213-219.
[207] Brusic, V., Rudy, G., Kyne, A.P. and Harrison, L.C. (**1996**) *Nucleic Acids Res.*, *24*, 242-244.
[208] Blythe, M.J., Doytchinova, I.A. and Flower, D.R. (**2002**) *Bioinformatics*, *18*, 434-439.
[209] Bhasin, M., Singh, H. and Raghava, G.P. (**2003**) *Bioinformatics*, *19*, 665-666.
[210] Schonbach, C., Koh, J.L., Sheng, X., Wong, L. and Brusic, V. (**2000**) *Nucleic Acids Res.*, *28*, 222-224.
[211] Holm, L. and Sander, C. (**1996**) *Science*, *273*, 595-603.
[212] Veropoulos, K., Campbell, C. and Cristianini, N. (**1999**) Controlling the sensitivity of Support Vector machines. In Proceedings of the International Joint Conference on Artificial Intelligence (UCAI99), (Dean, T. ed.). pp. 55-60. Sweden: Morgan Kaufmann.
[213] Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr. and Haussler, D. (**2000**) *Proc. Natl. Acad. Sci. USA*, *97*, 262-267.
[214] Furlanello, C., Serafini, M., Merler, S. and Jurman, G. (**2003**) *Neural Networks*, *16*, 641-648.
[215] Xue, Y., Yap, C., Sun, L., Cao, Z., Wang, J. and Chen, Y. (**2004**) *J. Chem. Infor. Sci.*, *44*, 1497-1505.
[216] Li, H., Ung, C., Yap, C., Xue, Y., Li, Z., Cao, Z. and Chen, Y. (**2005**) *Chem. Res. Toxicol.*, *18*, 1071-1080.
[217] Yap, C.W. and Chen, Y.Z. (**2005**) *J. Chem. Infor. Model.*, *45*, 982-992.
[218] Deshmukh, S., Khaitan, S., Das, D., Gupta, M. and Wangikar, P.P. (**2007**) *Protein Pept. Lett.*, *14*, 647-657.
[219] Zhou, X.B., Chen, C., Li, Z.C. and Zou, X.Y. (**2007**) *J. Theor. Biol.*, *248*, 546-551.
[220] Mishra, N.K., Kumar, M. and Raghava, G.P. (**2007**) *Protein Pept. Lett.*, *14*, 575-580.
[221] Lin, H.H., Han, L.Y., Cai, C.Z., Ji, Z.L. and Chen, Y.Z. (**2006**) *Proteins*, *62*, 218-231.
[222] Saha, S., Zack, J., Singh, B. and Raghava, G.P. (**2006**) *Genomics Proteomics Bioinformatics*, *4*, 253-258.
[223] Strope, P.K. and Moriyama, E.N. (**2007**) *Genomics*, *89*, 602-612.
[224] Fujishima, K., Komasa, M., Kitamura, S., Suzuki, H., Tomita, M. and Kanai, A. (**2007**) *DNA Res.*, *14*, 91-102.
[225] Lin, H.H., Han, L.Y., Zhang, H.L., Zheng, C.J., Xie, B. and Chen, Y.Z. (**2006**) Submitted.
[226] Bhardwaj, N. and Lu, H. (**2007**) *FEBS Lett.*, *581*, 1058-1066.
[227] Fang, Y., Guo, Y., Feng, Y. and Li, M. (**2007**) *Amino Acids*.
[228] Lin, H.H., Han, L.Y., Zhang, H.L., Zheng, C.J., Xie, B., Cao, Z.W. and Chen, Y.Z. (**2006**) *BMC Bioinformatics*, *7 Suppl 5*, S13.
[229] Cai, Y.D., Zhou, G.P. and Chou, K.C. (**2003**) *Biophys. J.*, *84*, 3257-3263.
[230] Wang, M., Yang, J., Liu, G.P., Xu, Z.J. and Chou, K.C. (**2004**) *Protein Eng. Des. Sel.*, *17*, 509-516.
[231] Park, K.J., Gromiha, M.M., Horton, P. and Suwa, M. (**2005**) *Bioinformatics*, *21*, 4223-4229.
[232] Tan, F., Feng, X., Fang, Z., Li, M., Guo, Y. and Jiang, L. (**2007**) *Amino Acids*.
[233] Xu, H., Xu, H., Lin, M., Wang, W., Li, Z., Huang, J., Chen, Y. and Chen, X. (**2007**) *Proteomics*.
[234] Chen, X., Fang, Y., Yao, L., Chen, Y. and Xu, H. (**2007**) *Methods Inf. Med.*, *46*, 360-366.
[235] Li, Q. and Lai, L. (**2007**) *BMC Bioinformatics*, *8*, 353.
[236] Saha, S. and Raghava, G.P. (**2006**) *Nucleic Acids Res.*, *34*, W202-209.
[237] Cui, J., Han, L.Y., Li, H., Ung, C.Y., Tang, Z.Q., Zheng, C.J., Cao, Z.W. and Chen, Y.Z. (**2007**) *Mol. Immunol.*, *44*, 514-520.
[238] Soeria-Atmadja, D., Lundell, T., Gustafsson, M.G. and Hammerling, U. (**2006**) *Nucleic Acids Res.*, *34*, 3779-3793.
[239] Zhang, G. and Fang, B. (**2006**) *Protein Pept. Lett.*, *13*, 965-970.
[240] Mitra, J., Mundra, P., Kulkarni, B.D. and Jayaraman, V.K. (**2007**) *J. Biomol. Struct. Dyn.*, *25*, 289-298.
[241] Ansari, F.A., Kumar, N., Bala Subramanyam, M., Gnanamani, M. and Ramachandran, S. (**2007**) *Proteins* [Epubahead of print]
[242] Lata, S., Sharma, B.K. and Raghava, G.P. (**2007**) *BMC Bioinformatics*, *8*, 263.
[243] Mondal, S., Bhavna, R., Mohan Babu, R. and Ramakumar, S. (**2006**) *J. Theor. Biol.*, *243*, 252-260.
[244] Zhao, Y., Pinilla, C., Valmori, D., Martin, R. and Simon, R. (**2003**) *Bioinformatics*, *19*, 1978-1984.
[245] Donnes, P. and Elofsson, A. (**2002**) *BMC Bioinformatics*, *3*, 25.
[246] Cui, J., Han, L.Y., Lin, H.H., Tang, Z.Q., Jiang, L., Cao, Z.W. and Chen, Y.Z. (**2006**) *Immunogenetics*, *58*, 607-613.
[247] Cui, J., Han, L.Y., Lin, H.H., Zhang, H.L., Tang, Z.Q., Zheng, C.J., Cao, Z.W. and Chen, Y.Z. (**2007**) *Mol. Immunol.*, *44*, 866-877.
[248] Zhang, G.L., Khan, A.M., Srinivasan, K.N., August, J.T. and Brusic, V. (**2005**) *Nucleic Acids Res.*, *33*, W172-179.
[249] Zhang, G.L., Bozic, I., Kwoh, C.K., August, J.T. and Brusic, V. (**2007**) *J. Immunol. Methods*, *320*, 143-154.
[250] Bhasin, M. and Raghava, G.P. (**2004**) *Bioinformatics*, *20*, 421-423.
[251] Yang, Z.R. and Johnson, F.C. (**2005**) *J. Chem. Inf .Model.*, *45*, 1424-1428.
[252] Han, L.Y., Zheng, C.J., Lin, H.H., Cui, J., Li, H., Zhang, H.L., Tang, Z.Q. and Chen, Y.Z. (**2005**) *New Phytol.*, *168*, 109-121.

Received: November 18, 2006 Revised: June 19, 2007 Accepted: June 19, 2007