REVIEW

Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity

Lianyi Han[1], Juan Cui[1], Honghuang Lin[1], Zhiliang Ji[2], Zhiwei Cao[2], Yixue Li[2] and Yuzong Chen[1, 2]

[1] Department of Computational Science and Department of Pharmacy, National University of Singapore, Singapore, Singapore
[2] Shanghai Center for Bioinformation Technology, Shanghai, P. R. China

Protein sequence contains clues to its function. Functional prediction from sequence presents a challenge particularly for proteins that have low or no sequence similarity to proteins of known function. Recently, machine learning methods have been explored for predicting functional class of proteins from sequence-derived properties independent of sequence similarity, which showed promising potential for low- and non-homologous proteins. These methods can thus be explored as potential tools to complement alignment- and clustering-based methods for predicting protein function. This article reviews the strategies, current progresses, and underlying difficulties in using machine learning methods for predicting the functional class of proteins. The relevant software and web-servers are described. The reported prediction performances in the application of these methods are also presented, which need to be interpreted with caution as they are dependent on such factors as datasets used and choice of parameters.

Received: January 4, 2006
Revised: March 23, 2006
Accepted: March 25, 2006

Keywords:
Machine learning method / Neural network / Protein function prediction / Protein sequence / Support vector machine

1 Introduction

Protein sequence contains clues to its function [1, 2]. These clues have been explored for predicting protein function using sequence similarity [3–5], clustering [6–8], motifs [9], and evolutionary relationships [10, 11]. These methods tend to become less effective for proteins that lack clear sequence or structural similarity to proteins of known function [10, 12, 13]. As these "low homologous" and "non-homologous" proteins constitute a substantial percentage, up to 20–100%, of the ORFs in many of the currently completed genomes [14], there is a need for exploring functional prediction methods independent of sequence similarity [2, 15]. Moreover, not all homologous proteins have analogous functions [11]. The presence of shared domain within a group of proteins does not necessarily imply that these proteins perform the same function [16]. Many proteins sharing promiscuous domains are known to have very different functions [17]. 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 [18]. Hence, methods independent of sequence similarity are also needed to complement sequence similarity and clustering methods for reliable prediction of the function of these proteins [2, 15].

Correspondence: Professor Y. Z. Chen, Department of Computational Science and Department of Pharmacy, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, Singapore
E-mail: yzchen@cz3.nus.edu.sg
Fax: +65-6774-6756

Abbreviations: ANN, artificial neural network; **GPCR**, G-protein coupled receptor; **SVM**, support vector machine

Significant interest and efforts have been directed at the development of such similarity-independent methods. So far, two approaches have been primarily explored in most studies. One predicts protein function using such non-sequence information or biological hypothesis as structural features [19, 20], interaction profiles [21, 22], and protein/gene fusion data [17, 23]. Another applies machine learning methods [24–34] for predicting protein functional classes from sequence-derived physicochemical properties [22, 35–37], sequence-generated signatures [26, 36, 38, 39], PTMs and localization features [25, 40, 41], and combination of sequence features with other profiles such as function/network features [42], and for generating prediction rules based on combination evidence from amino acid attributes, predicted structure and phylogenic pattern [43, 44]. In particular, machine learning methods have shown promising potential for predicting the functional class of proteins irrespective of sequence similarity [14, 27, 29, 43–46].

Machine learning methods have been applied to the prediction of proteins of specific functional class characterized by distinguished biochemical properties or biological activities. These classes include G-protein coupled receptors [26, 47, 48], nuclear receptors [31], transmembrane proteins [49, 50], DNA-binding proteins [28, 51], RNA-binding proteins [28, 30], lipid-binding proteins [52], enzymes of various families [29, 32, 33, 53], and transporters of various families [54]. These methods have also been applied to the prediction of broadly defined classes of proteins with multiple biochemical properties or biological activities. These broadly defined classes include cytokines [55], hormone proteins [34], stress response proteins [34], receptors [34], crystallizable proteins [56], mitochondrial proteins [57], cell cycle-regulated proteins in yeast [41, 58], and functional classes in yeast [59]. While most methods are yet to be employed in real case studies, one method has recently been successfully applied to the study of a practical biological problem, the prediction of cell cycle phases as well as cell cycle-regulated genes [58, 60].

Proteins in each of these broadly defined functional classes may belong to multiple functional families but, nonetheless, share some common characteristics. For instance, stress response proteins have a strong bias for β-sheets especially in C-terminal regions, and they usually contain signal peptides [34]. Crystallizable proteins tend to be shorter in length and lower in hydrophobicity, have fewer interacting partners, and contain serine and charged residues, which are essential for the expression, solubilization, purification, and crystallization of these proteins [56]. These common characteristics may be exploited to distinguish members and non-members of each class by machine learning methods.

Machine learning methods have also shown some level of capability in predicting the functional class of such novel proteins as remote homologs, homologous proteins of different functions, and proteins non-homologous to any protein of known function [14, 45, 46, 61, 62]. This capability has been attributed to several factors. These methods are capable of accommodating low similarity proteins [61]. In these methods, proteins are represented by some form of descriptors instead of straightforward use of sequence, and these descriptors either describe physicochemical properties of the constituent amino acids [14, 21, 45] or capture both local structural motifs and longer conserved regions associated with specific functional properties [62] of the protein they represent. Many types of functionalities require the presence and recognition of short sequence motifs for PTM or binding of other factors to the protein, but do not require the sequence or structure to be conserved [41].

This article reviews the strategies, current progresses and underlying difficulties in the application of machine learning methods for predicting the functional class of proteins. The reported prediction performances in the application of these methods are also described, which needs to be interpreted with caution as these prediction results are dependent on such factors as the datasets used and choice of parameters. Proper representation of proteins, particularly their structural and physicochemical properties, is one of the keys to the successful application of machine learning methods. A variety of protein descriptors [22, 26, 27, 63–66] have been derived to quantitatively represent various structural and physicochemical properties of proteins. The algorithms for deriving these descriptors are discussed. This article also describes the web-servers for computing these descriptors and for predicting protein functional classes using machine learning methods.

2 Strategies for using machine learning methods to predict functional class of proteins

One strategy for predicting the function of a protein without using sequence similarity or clustering is to predict the functional class of that protein using a sequence-independent classifier or a set of rules generated from the analysis of a sufficiently diverse set of proteins that share the common functional characteristics, but may be significantly different in sequence and structure [28–30, 53, 59]. The advantage of this approach is that more information can be extracted from multiple proteins in a functional class that share these common functional characteristics, which can be used to derive sequence profiles [6–8] and classifiers [22, 28–30, 35–37, 42, 53, 59, 67–70] for predicting other proteins that have the same characteristics. Known members and non-members of a functional class form two separate groups. Thus, the two-class classification machine learning methods can be applied for developing an artificial intelligence system to separate members and non-members. Examples of two-class systems are members and nonmembers of G-protein coupled receptors (GPCRs) [47, 48]. Some proteins, such as members of sub-families of GPCRs and enzymes, can be organized into multiple functional classes at the sub-family levels. Multi-class classification methods can be used for developing functional prediction systems for these proteins [26, 53, 55].

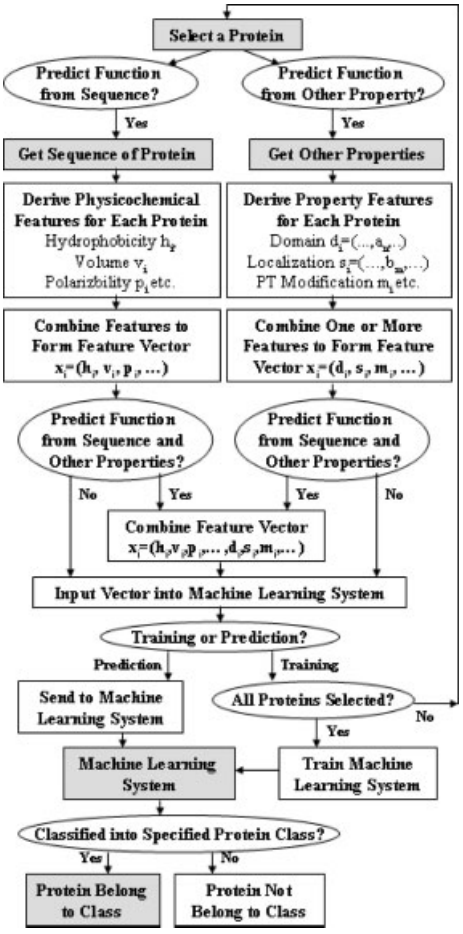

Figure 1. Schematic diagram illustrating the process of the training and prediction of the functional class of proteins from sequence or other properties using a machine learning method. Sequence-derived feature h_j, p_j, vj . . . represents such structural and physicochemical properties as hydrophobicity, polarizability, and volume. Feature d_j, s_j, mj . . .represents such properties as domain information, subcellular localization, and PTM (PT) profiles, etc.

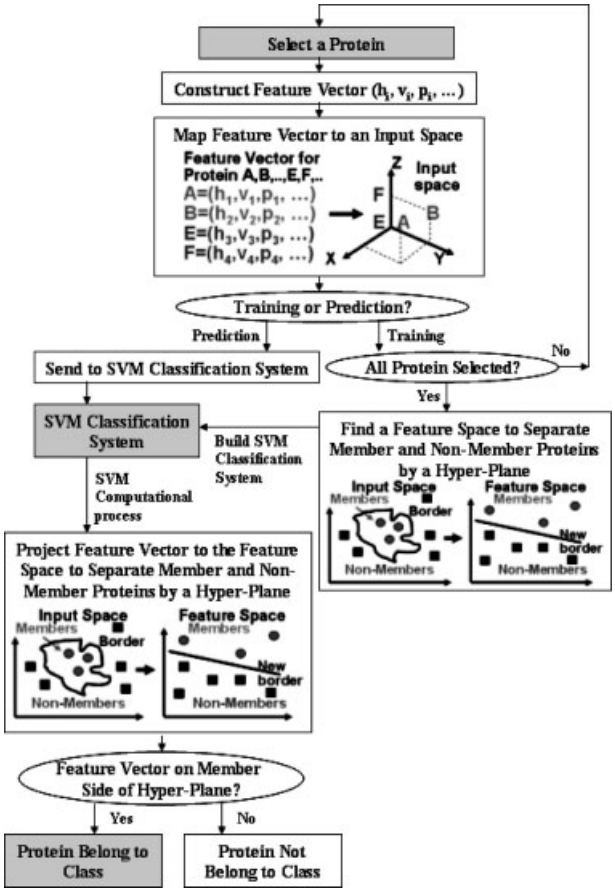

Figure 2. Schematic diagram illustrating the process of the training and prediction of the functional class of proteins using the support vector machine (SVM) method. (A, B) Feature vectors of proteins belong to a functional class; (E, F) feature vectors of proteins not belong to a functional class. h_j, p_j, vj, are the same as in Fig. 1.

A developed machine learning classification system can be subsequently used for classifying a new protein into one of the two or multiple classes, and it can be predicted to have the same functional profile if it is classified as a member. Sequence-derived features have frequently been used for representing proteins [22, 28–30, 35–37, 42, 53, 59, 67–69] in the development of the classification systems for predicting the functional class of proteins.

Figure 1 illustrates the process of using a machine learning method for training and predicting proteins that have a specific common functional profile. Protein members and non-members of a functional class are represented by separate sets of feature vectors, which are composed of descriptors derived either from the sequence of these proteins for representing their structural and physicochemical properties or from other properties such as PTM and subcellular localization. A machine learning classification system can be trained to separate these two sets of feature vectors into separate classes, and a new protein can be predicted to be a member or non-member of the functional class if its feature vector is classified into the respective class by this machine learning classification system.

Figures 2 and 3 illustrate the training and prediction process of two specific methods, support vector machines (SVM) and artificial neural networks (ANN). SVM classifies proteins by projecting their feature vectors into a multi-dimensional space in which members and non-members of a functional class are separated by a hyper-plane. A new protein can be predicted to be a member or non-member if its feature vector is projected on the side of the hyper-plane where other proteins having the same profile are located. ANN uses a similar procedure but different machine learning algorithm for training a two-class classification system and for using that system to predict protein functional class.

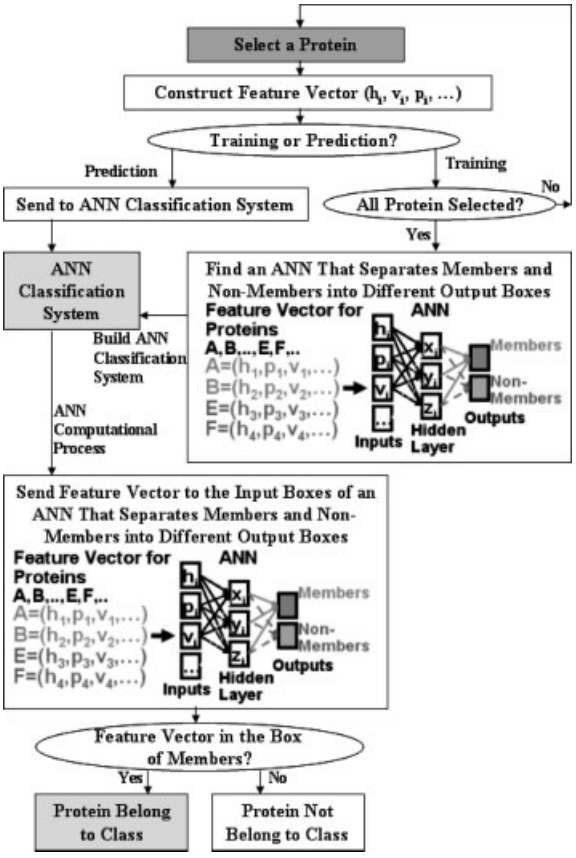

Figure 3. Schematic diagram illustrating the process of the prediction of functional class of proteins using the machine learning method, ANN. (A, B, E, F) h_j, p_j, vj, are the same as in Figs. 1 and 2.

3 Representation of protein sequences

A number of descriptors have been introduced for representing protein sequence [22, 26, 27, 63–67], PTMs and localization features [25, 40, 41]. The sequence-derived descriptors include amino acid composition, dipeptide composition, sequence autocorrelation descriptors, sequence coupling descriptors, and the descriptors for the composition, transition and distribution of hydrophobicity, polarity, polarizibility, charge, secondary structures, and normalized Van der Waals volumes. Web servers such as PROFEAT (http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi) and ProtParam [66] (http://www.expasy.org/tools/protparam.html) have also appeared for facilitating the computation of these descriptors. Other sequence derived features such as cleavage sites, nuclear export signals, and subcellular localization can be computed from such web servers as CBS Prediction Servers (http://www.cbs.dtu.dk/services/).

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 [31]. Autocorrelation descriptors are defined from the distribution of amino acid properties along the sequence [71]. The amino acid indices used in these autocorrelation descriptors include hydrophobicity scales [72], average flexibility indices [73], polarizability parameter [74], free energy of solution in water [74], residue accessible surface area in proteins [75], residue volume [76], steric parameter [77], and relative mutability [78]. Each of these indices is centralized and standardized before the calculation. Moreau-Broto autocorrelation descriptors [79] are defined as

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \tag{1}$$

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 *ATS(d) = AC(d)/(N − d), where d = 1, 2, 3, . . ., 30. Moran autocorrelation descriptors [80] are defined as:*

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}\left(P_i - \overline{P}\right)\left(P_{i+d} - \overline{P}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2} \tag{2}$$

for d = 1, 2, 3, . . ., 30, and 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 [81] are defined as:

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}\left(P_i - P_{i+d}\right)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2} \tag{3}$$

for d = 1, 2, 3, . . ., 30, and 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 [63–65] and the Grantham chemical distance matrix [82] between the 20 amino acids. The d-th-rank sequence-order-coupling number is defined as

$$\tau_d = \sum_{i=1}^{N-d}\left(d_{i,i+d}\right)^2 \tag{4}$$

for 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:

$$Xr = \frac{f_r}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{30}\tau_d} \tag{5}$$

for 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:

$$Xd = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{30} \tau_d} \quad (6)$$

for d = 21, 22, 23, . . ., 50.

Three descriptors, composition (C), transition (T) and distribution (D), are derived for each of the following physicochemical properties: hydrophobicity, polarity, polarizibility, charge, secondary structures, and normalized Van der Waals volume [27, 83, 84]. For every property, the constituent amino acids in a protein 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. C is the number of amino acids in a particular class divided by the total number of amino acids. T characterizes the percent frequency with which amino acids of a particular class is followed by amino acids of a different class. D measures the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular class is located, respectively.

4 Algorithms of machine learning methods

Several machine learning methods have been widely used for the classification of proteins. These include ANN, Probabilistic Neural Network (PNN), Decision Tree (DT) and SVM. Websites for the freely downloadable codes of these methods are given in Table 1. There are also web-servers, listed in Table 2, that allow users to predict the functional class of their own protein using some of these machine learning methods.

4.1 Artificial Neural Network

ANN is a machine learning method inspired by the biological nervous system, which trains a hidden-layer-containing network and uses its connected structures for pattern recognition and classification [85, 86]. A classifier for ANN is usually in the form of

$$y = g\sum_j w_0 h_j \quad (7)$$

where w_{0j} is the output weight of a hidden node j to an output node, g is the output function, h_j is the value of a hidden layer node:

$$h_j = \delta\left(\sum_j w_{ji}x_j + w_j\right) \quad (8)$$

x_i represents the feature vector of a protein whose components are their computed descriptors, w_{ji} is the input weight from an input node i to a hidden node j, w_j is the threshold weight from an input node of value 1 to a hidden node j, and δ is an active function where the sigmoid function is mostly used.

4.2 Probabilistic Neural Network

PNN is a form of neural network that uses Bayes optimal decision rule for classification [87]. Traditional neural networks such as feed-forward back-propagation neural network rely on multiple parameters and network architectures to be optimized. In contrast, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's non-parameteric estimator [88]. Thus, the training process of PNN is usually orders of magnitude faster than those of the traditional neural networks.

4.3 k Nearest Neighbor

In k Nearest Neighbor (kNN), the Euclidean distance between an unclassified feature vector $\boldsymbol{x}$ and each individual feature vector $\boldsymbol{x}_i$ in the training set is measured [89, 90]. A total of k number of vectors nearest to the unclassified vector $\boldsymbol{x}$ are used to determine the class of that unclassified vector. The class of the majority of the k nearest neighbors is chosen as the predicted class of the unclassified vector $\boldsymbol{x}$.

4.4 Decision Tree

DT is a branch-test-based classifier [91]. A branch of the DT corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test on a single attribute value, with one branch and its subsequent classes as possible outcomes. C4.5 DT uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A feature vector $\boldsymbol{x}$ is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted to move into a branch. Upon reaching the destination leaf, the class of the vector $\boldsymbol{x}$ is predicted to be that of the leaf.

4.5 Support Vector Machine

There are two types of SVM algorithms, linear and nonlinear SVM. Nonlinear SVM is more useful for classifying proteins of diverse sequences and small molecules of diverse structures, and it has thus been more extensively used. Linear SVM constructs a hyper-plane separating two different classes of feature vectors with a maximum margin [92, 93]. This hyper-plane is constructed by finding a vector $\mathbf{w}$ and a parameter b that minimizes $\|\mathbf{w}\|^2$ which satisfies the following conditions: $\mathbf{w}\cdot\mathbf{x}_i + b \geq +1$, for $y_i = +1$ (positive class) and $\mathbf{w}\cdot\mathbf{x}_i + b \leq -1$, for $y_i = -1$ (negative class). Here $\mathbf{x}_i$ is a fea-

Table 1. Websites that contain freely downloadable codes of machine learning methods

Decision Tree	
PrecisionTree	http://www.palisade.com.au/precisiontree/
DecisionPro	http://www.vanguardsw.com/decisionpro/jdtree.htm
C4.5	http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html
C5.0	http://www.rulequest.com/download.html
KNN	
k Nearest Neighbor demo	http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html
PERL Module for KNN	http://aspn.activestate.com/ASPN/CodeDoc/AI-Categorize/AI/Categorize/kNN.html
Java class for KNN	http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/classify/old/KNN.html
Neural Network	
BrainMaker	http://www.calsci.com/
Libneural	http://pcrochat.online.fr/webus/tutorial/BPN_tutorial7.html
fann	http://leenissen.dk/fann/
NeuralWorks Predict	http://www.neuralware.com/products.jsp
NeuroShell Predictor	http://www.mbaware.com/neurpred.html
SVM	
SVM light	http://svmlight.joachims.org/
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
SMO	http://www.datalab.uci.edu/people/xge/svm/
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/

Table 2. Web servers for computing protein functional classes using machine learning methods

Web-server	URL
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
ProtFun: ANN prediction of cellular role, enzyme class and Gene Ontology category	http://www.cbs.dtu.dk/services/ProtFun/
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

ture vector, y_i is the group index, $\mathbf{w}$ is a vector normal to the hyper-plane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of $\mathbf{w}$. Nonlinear SVM projects feature vectors into a high dimensional feature space using a kernel function such as $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_i\|^2/2\sigma^2}$. The linear SVM procedure is then applied to the feature vectors in this feature space. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}$ can be classified using $sign[(\mathbf{w}\cdot\mathbf{x}) + b]$, a positive or negative value indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

4.6 Performance measurement

The performance of machine learning methods has been measured using the positive prediction accuracy $P_+ = TP/(TP+FN)$ for proteins that have a specific property and the negative prediction accuracy $P_- = TN/(TN+FP)$ for proteins without that property [22, 28–30, 35–37, 42, 53, 59, 67–70]. 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 molecules, 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 [94, 95]. Thus, the Matthews correlation coefficient

$$\mathrm{MCC} = (\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN})/\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}$$

has been used for measuring the performance of machine learning methods [29–31, 47, 55, 57].

5 Assessment of the performance of machine learning methods

5.1 Performance for predicting protein functional classes

Table 3 summarizes the reported performance of the use of machine learning methods for predicting protein functional classes. These were selected from relevant publications that

Table 3. 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. The reported prediction performances need to be interpreted with caution as they are dependent on such factors as the datasets used and choice of parameters

Protein functional class	Protein sub-classes	Method	Protein descriptors	Number of proteins in training set N (N+/N−)	Validation method	Reported prediction accuracy				Ref.
						P+ (%)	P− (%)	P (%)	MCC	
G-protein coupled receptors	All GPCRs	SVM	Physicochemical properties	2247 (927/1320)	Independent evaluation	95.6	98.1	97.4	0.93	[27]
		SVM	Dipeptide composition	3302 (778/2524)	5-fold CV	98.6	99.8	99.5	0.99	[47]
	Gi/o binding type	SVM	Structural characteristics (extra cellular loops, intracellular loops etc)	132 (61/71)	4-fold CV	77.0	78.3			[48]
	Gq/11 binding type	SVM		132 (47/85)	4-fold CV	68.1	72.7			
	Gs binding type	SVM		132 (24/108)	4-fold CV	83.3	95.2			
Nuclear receptors		SVM	Amino acid composition	282	5-fold CV			82.6	0.74	[31]
			Dipeptide composition	282	5-fold CV			7.5	0.96	
			Physicochemical properties	872 (334/538)	Independent evaluation	89.5	97.6			[27]
Receptors		ANN	Post translational modifications, protein sorting signals and physical/chemical properties calculated from the amino acid composition			∼85	∼90			[34]
Enzymes	All enzymes	ANN	Post-translational modifications and localization features	4658 (1620/4038)	Independent evaluation	∼75	∼75			[25]
	EC1 Oxidoreductase			1532 (319/1213)		∼70	∼70			
	EC2 Transferase			1532 (529/1003)		∼80	∼70			
	EC3 Hydrolase			1532 (485/1047)		∼70	∼65			
	EC4 Lyase			1532 (78/1454)		∼80	∼70			
	EC5 Isomerase			1532 (72/1460)		∼75	∼85			
	EC6 Ligase			1532 (49/1483)		∼87	∼90			
	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	SVM	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	[27, 29]
	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	SVM	Functional domain composition and pseudo amino acid composition	503∼3582 (3∼2002/327∼3548)	Jackknife test	25.0∼100.0				[32]

Table 3. Continued

Protein functional class	Protein sub-classes	Method	Protein descriptors	Number of proteins in training set N (N+/N−)	Validation method	Reported prediction accuracy				Ref.
						P+ (%)	P− (%)	P (%)	MCC	
Transporters	Transporter	ANN	PTMs, protein sorting signals and physical/chemical properties calculated from the amino acid composition	N.A	5-fold CV	~75	~80			[34]
	Ion channel			N.A		~70	~75			
	Voltage-gated ion channel			N.A		~75	~77			
	Cation channel			N.A		~60	~80			
	Metal ion transporter			N.A		~60	~65			
	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	SVM	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	[54]
RNA-binding proteins	All RNA-binding proteins	SVM	Amino acid composition and limited range correlation of hydrophobicity and solvent accessible surface area	6264 (1496/4768)	10-fold CV	76.5	97.2	92.2		[28]
		SVM	Physicochemical properties	5126 (2161/2965)	Independent evaluation	97.8	96.0	96.1	0.80	[30]
	rRNA-binding	SVM	Amino acid composition, limited range correlation of hydrophobicity, solvent accessible surface area	5824 (1056/4768)	10-fold CV	100	99.9	99.9		[28]
		SVM	Physicochemical properties	1680 (708/972)	Independent evaluation	94.1	98.7	98.6	0.74	[30]
	tRNA-binding	SVM	Physicochemical properties	886 (94/792)	Independent evaluation	94.1	99.9	99.8	0.92	[30]
	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	SVM	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]
		SVM	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		[51]
					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		
		SVM	Physicochemical properties	8168 (4587/3581)	Independent evaluation	85.7	93.9	91.2	0.80	[27]

Table 3. Continued

Protein functional class	Protein sub-classes	Method	Protein descriptors	Number of proteins in training set N (N+/N−)	Validation method	Reported prediction accuracy				Ref.
						P+ (%)	P− (%)	P (%)	MCC	
	DNA condensation	SVM	Physicochemical properties	965 (128/837)	Independent evaluation	89.2	99.7	99.5	0.85	[27]
	DNA integration			1595 (239/1356)		89.1	99.9	99.6	0.91	
	DNA recombination			3582 (881/2701)		87.0	98.9	97.7	0.87	
	DNA repair			5086 (1343/3743)		88.8	97.2	95.8	0.85	
	DNA replication			4938 (1520/3418)		78.9	98.7	96.5	0.82	
	Transcription factors			1878 (426/1452)		82.9	99.7	99.0	0.87	
Lipid-binding proteins	All lipid-binding proteins	SVM	Physicochemical properties	6933 (3232/3701)	Independent evaluation	89.9	97.0	94.1	0.88	[27, 52]
	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	
Transmembrane proteins		SVM	Functional domain composition	2059	Jackknife			86.3		[49]
					Independent			67.5		
					self-consistency			93.9		
		SVM	Pseudo-amino acid composition	2059	Jackknife			82.4		[50]
					Independent			90.3		
					self-consistency			99.9		
		SVM	Physicochemical properties	4668 (2105/2563)	Independent evaluation	90.1	86.7	86.7	0.75	[27]
Cytokines	All cytokines	SVM	Dipeptide composition	1110 (437/673)	7-fold CV	92.5	97.2	95.3	0.90	[55]
	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		6.7∼100	85.5∼100	84∼98	0.65∼0.96	
Transcription		ANN	PTMs, protein sorting signals and physical/chemical properties calculated from the amino acid composition		5-fold CV	∼75	∼70			[34]
Transcription regulation						∼80	∼75			
Growth factor						∼65	∼75			
Crystallizable proteins		SVM + Naive Bayes classifier	Mono-, di-, tri-peptide composition, physicochemical and structural properties	923 (721/202)	10-fold CV	65.0	69.0	67.0		[56]

Table 3. Continued

Protein functional class	Protein sub-classes	Method	Protein descriptors	Number of proteins in training set N (N+/N−)	Validation method	Reported prediction accuracy				Ref.
						P+ (%)	P− (%)	P (%)	MCC	
Hormones		ANN	PTMs, protein sorting signals and physical/chemical properties calculated from the amino acid composition	N.A	5-fold CV	~79	~90			[34]
Stress response				N.A (65/N.A)		~77	~90			
Immune response				N.A (48/N.A)		~80	~80			
Signal transducer				N.A (462/N.A)		~75	~78			
Structural proteins				N.A		~80	~77			
Mitochondrial proteins		SVM + HMM	Amino acid composition	10372 (1432/8940)	5-fold CV	78.9	90.0	88.2	0.62	[57]
Cell cycle regulated proteins		ANN	PTMs and localization features	671 (115/556)	3-fold CV	37.8	96.2			[41]
Functional classes in Mycobacterium tuberculosis	Small molecule metabolism (degradation, energy metabolism, etc.), macromolecule metabolism, cell processes, others	DMP (logic programming data mining + decision tree)	Amino acid attributes, predicted structure and phylogenic patterns	3924	Independent set	62~76				[44]
Functional classes in E. coli	Small molecule metabolism (degradation, energy metabolism, etc.), macromolecule metabolism, cell processes, others	DMP (logic programming data mining + decision tree)	Amino acid attributes, predicted structure and phylogenic patterns	4289	Independent set	75~100				[43]
Functional classes in yeast	All proteins 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	kNN + SVM	Functional domain composition	4902 86~725	Jackknife Jackknife			72.0 15~90		[59]

provide sufficient information about dataset, machine learning method, and protein functional class prediction accuracy. All of the data and results shown in Table 3 are from the original papers. The reported P_+ and P_- values are in the range of 65–100% and 69.0–99.9%, with the majority concentrated in the range of 75–95% and 95–99.9%, respectively. Based on these reported results, machine learning methods generally show certain level of capability for predicting the functional class of proteins. 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 a more diverse set of non-members than that of members, which enables machine learning methods to perform a better machine learning for recognition of non-members. It needs to be pointed out that the reported prediction performances are sensitively dependent on such factors as the sizes and diversity of protein samples used, and the choice of parameters of the machine leaning methods. Therefore, caution needs to be exercised when comparing and interpreting these results.

5.2 Performance for predicting functional profile of novel proteins

The performance of machine learning methods for predicting the functional profile of novel proteins has also been evaluated in several studies [14, 45, 46, 61]. The studied 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 [2, 15]. Machine learning methods have been reported to have certain level of capacity for predicting the functional class of various types of novel proteins. Examples of the reported prediction accuracies are 46% of the remote homologs from FSSP database [61], 66–76% of the proteins of known function without a single homolog in NR or the Swiss-Prot database [14, 45, 46], and 62% of the pairs of homologous enzymes of different families found from the Swiss-Prot database [14].

The ability of machine learning methods in predicting the functional profile of the first type of novel proteins have been attributed to the non-discriminative nature of such machine learning methods as SVM for selecting class members, and to the use of structural and physicochemical descriptors for representing proteins [14, 45, 46, 61, 62]. 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 of the same functional property [96]. 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 [97, 98], 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 [20].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.

5.3 Underlying difficulties in the application of machine learning methods

The performance of machine learning methods critically depends on the diversity of samples (proteins) 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 protein members and non-members of a particular functional class. Various degrees of inadequate sampling representation likely affect, to a certain extent, the prediction accuracy of the developed machine learning models. Machine learning methods are not applicable for proteins with insufficient knowledge about their specific functional profile. Searching of the information about proteins known to possess a particular profile and those do not possess that profile is a key to more extensive exploration of machine learning methods for facilitating the study of protein functional profiles. Apart from literature sources such as PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed), a number of databases are useful for obtaining information about protein functional profiles, which are listed in Table 4.

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. Some machine learning methods are known to produce biased results for an imbalanced datasets. For instance, the SVM method tends to produce feature vectors that push the hyper-plane towards the side with a smaller number of data [99], 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 re-adjusting biased shift of hyper-plane are being explored [100]. Applica-

Table 4. Useful databases for obtaining information about protein functional profiles

Database	Information	URL	Ref.
PubMed	Literatures	http://www.ncbi.nih.gov/	[112]
UniProt Knowledgebase	Functional information of proteins	http://www.expasy.uniprot.org/database/knowledgebase.shtml	[113]
PIR-PSD	Classified and functionally annotated protein sequences	http://pir.georgetown.edu/pirwww/dbinfo/pir_psd.shtml	[114]
GenBank	DNA sequence	http://www.ncbi.nlm.nih.gov/Genbank/	[115]
Gene Ontology	Description of gene and gene product attributes	http://www.geneontology.org/	[116]
KEGG	Encyclopedia of genes, genomes and pathways	http://www.genome.ad.jp/kegg/kegg2.html	[117]
PDB	Protein 3-D structure	http://www.rcsb.org/pdb/Welcome.do	[109]
SCOP	Structural classification of proteins	http://scop.mrc-lmb.cam.ac.uk/scop/	[118]
FSSP database	Families of structurally similar proteins	http://swift.cmbi.kun.nl/swift/fssp/	[119]
PRINTS	Group of conserved motifs used to characterize a protein family	http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/	[120]
ENZYME database	Enzyme nomenclature	http://www.expasy.ch/enzyme/	[121]
BRENDA	Comprehensive enzyme information	http://www.brenda.uni-koeln.de/	[122]
TCDB	IUBMB approved classification system for membrane transport proteins	http://www.tcdb.org/	[123]
TransportDB	Predicted cytoplasmic membrane transport protein complement	http://www.membranetransport.org/	[124]
HMTD	Human membrane transporters	http://lab.digibench.net/transporter/	[125]
GPCRDB	Information system for G protein-coupled receptors	http://www.gpcr.org/7tm/index.html	[126]
TTD	Therapeutic targets	http://bidd.nus.edu.sg/group/cjttd/ttd.asp	[127]
DIP	Interacting proteins	http://dip.doe-mbi.ucla.edu/	[128]
KDBI	Kinetic data of protein and other biomolecular interactions	http://bidd.nus.edu.sg/group/kdbi/kdbi.asp	[129]
PDSP K_i database	Protein-ligand binding affinity	http://kidb.cwru.edu/	[130]
KiBank	Protein-ligand binding affinity, structures of chemicals and target proteins	http://kibank.iis.u-tokyo.ac.jp/	[131]
CLiBE	Computed protein-ligand binding energies	http://xin.cz3.nus.edu.sg/group/clibe/clibe.asp	[132]

tion of these methods may help improving the prediction accuracy of machine learning methods in the cases involving imbalanced data.

While a number of descriptors have been introduced for representing proteins [22, 26, 27, 63–66] and small molecules [101, 102], 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 machine learning methods [103–105]. 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 using feature selection methods [103–105]. Effort been also directed at the improvement of the efficiency and speed of feature selection methods [106], 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. Investigations 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 [70, 107, 108]. These have prompted works for developing new descriptors [51].

6 Concluding remarks and perspectives

Machine learning methods have been explored as potential tools for predicting protein functional profiles independently of sequence similarity. A number of studies have consistently demonstrated the usefulness that these methods have in predicting protein functional classes. Because of their sequence similarity-independent nature, these methods may be potentially useful for studying proteins that cannot be confidently predicted by sequence similarity or clustering approaches, *i.e.*, proteins exhibiting low or no sequence similarity to any protein of known function. One particular application is the study of the substantial number of unknown ORFs in many of the currently completed genomes that have no similar protein of known function.

Different machine learning prediction systems have been developed using sequence-derived features [27, 28, 31, 47], structural information [28, 48], PTMs and subcellular localization features [25, 34], functional domain composition [32, 59], protein sorting signals [34], and their combinations [28]. Sequence information is usually more abundant than the other types of information used in the studies described in this article. For instance, there are 35 246 protein 3-D structure entries in PDB database [109] as of Feb 21, 2006, compared to 208 005 protein sequence entries in Swiss-Prot database release 49.1, and these PDB entries belong to only ~15% of the 8183 protein families in Pfam database [110] version 19.0. Approximately 51.6% of protein entries in Swiss-Prot database (release 49.1) have known subcellular localization. Therefore, sequence-derived features are applicable to substantially higher number of proteins than other features, and they are particularly useful for the large number of proteins that have no other types of information is available.

Even though the performance of the studies using sequence-derived information is much better than expected, considering that only sequence has been used, the currently developed prediction systems may not be good enough for annotation purposes [34]. It has been shown by a number of studies that the use of the other types of information, such as structural characteristics [48], and PTMs and localization features [41], can substantially improve the predication performance. Therefore, in cases in which they are available, the other types of information should be used for facilitating the functional study of proteins, as demonstrated in the successful prediction of cell cycle phases as well as cell cycle-regulated genes [58, 60].

The practical application range of each of the commonly used protein features for genome annotation can be estimated using the yeast genome data as an example. Based on the data from UniProt database (including Swiss-Prot and Tremble) and the dbPTM database [111] as of Feb 21, 2006, the number of yeast ORFs with available information about each of these features are 37 798, 25 220, 477, 525, and 7001 for sequence, domain, 3-D structure, PTMs, and subcellular localization, respectively. Therefore, approximately 100%, 66.7%, 1.3%, 1.4%, and 18.5% of the ORFs in the yeast genome can be potentially predicted by machine learning methods based on the use of sequence-derived features, domain information, structural characteristics, PTMs, and subcellular localization, respectively. With the rapid progress in structural biology and proteomics, the prediction capability of machine learning methods can be further enhanced with increasing availability of biological data and more extensive knowledge about structure, transcription, and post-transcriptional processing features that define the functional profiles of proteins.

Proper use of descriptors for representing proteins may also help further improving the performance of machine learning methods for predicting the functional class of proteins. New descriptors can be introduced for better representing certain types of features that correlate with novel functional profiles. 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 machine learning methods. These efforts will enable the development of machine learning methods into useful tools for facilitating the study of functional profiles of proteins to complement other well-established methods such as sequence similarity and clustering methods.

This work was supported in part by grants from Singapore ARF R-151-000-031-112, Shanghai Commission for Science and Technology (04QMX1450) and the 973 National Key Basic Research Program of China (2004CB720103).

7 References

[1] Bork, P., Dandekar, T., Diaz–Lazcoz, Y., Eisenhaber, F. *et al.*, *J. Mol. Biol.* 1998, *283*, 707–725.

[2] Eisenberg, D., Marcotte, E. M., Xenarios, I., Yeates, T. O., *Nature* 2000, *405*, 823–826.

[3] Bork, P., Koonin, E. V., *Nat. Genet.* 1998, *18*, 313–318.

[4] Schuler, G. D., *Methods Biochem. Anal.* 1998, *39*, 145–171.

[5] Baxevanis, A. D., *Methods Biochem. Anal.* 1998, *39*, 172–188.

[6] Fujiwara, Y, Asogawa, M., *NEC Res. Dev.* 2002, *43*, 238–241.

[7] Enright, A. J., Ouzounis, C. A., *Bioinformatics* 2000, *16*, 451–457.

[8] Enright, A. J., Van Dongen, S., Ouzounis, C. A., *Nucleic Acids Res.* 2002, *30*, 1575–1584.

[9] Hodges, H. C., Tsai, J. W., *FASB J.* 2002, *16*, A543–A543.

[10] Eisen, J. A., *Genome Res.* 1998, *8*, 163–167.

[11] Benner, S. A., Chamberlin, S. G., Liberles, D. A., Govindarajan, S., Knecht, L., *Res. Microbiol.* 2000, *151*, 97–106.

[12] Whisstock, J. C., Lesk, A. M., *Q. Rev. Biophys.* 2003, *36*, 307–340.

[13] Rost, B., *J. Mol. Biol.* 2002, *318*, 595–608.

[14] Han, L. Y., Cai, C. Z., Ji, Z. L., Cao, Z. W. *et al.*, *Nucleic Acids Res.* 2004, *32*, 6437–6444.

[15] Smith, T. F., Zhang, X., *Nat. Biotechnol.* 1997, *15*, 1222–1223.

[16] Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P. *et al.*, *Science* 1997, *278*, 609–614.

[17] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W. *et al.*, *Science* 1999, *285*, 751–753.

[18] Shah, I., Hunter, L., *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1997, *5*, 276–283.

[19] Teichmann, S. A., Murzin, A. G., Chothia, C., *Curr. Opin. Struct. Biol.* 2001, *11*, 354–363.

[20] Todd, A. E., Orengo, C. A., Thornton, J. M., *J. Mol. Biol.* 2001, *307*, 1113–1143.

[21] Aravind, L., *Genome Res.* 2000, *10*, 1074–1077.

[22] Bock, J. R., Gough, D. A., *Bioinformatics* 2001, *17*, 455–460.

[23] Enright, A. J., Iliopoulos, I., Kyrpides, N. C., Ouzounis, C. A., *Nature* 1999, *402*, 86–90.

[24] des Jardins, M., Karp, P. D., Krummenacker, M., Lee, T. J., Ouzounis, C. A., *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1997, *5*, 92–99.

[25] Jensen, L. J., Gupta, R., Blom, N., Devos, D. *et al.*, *J. Mol. Biol.* 2002, *319*, 1257–1265.

[26] Karchin, R., Karplus, K., Haussler, D., *Bioinformatics* 2002, *18*, 147–159.

[27] Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., Chen, Y. Z., *Nucleic Acids Res.* 2003, *31*, 3692–3697.

[28] Cai, Y. D., Lin, S. L., *Biochim. Biophys. Acta* 2003, *1648*, 127–133.

[29] Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, Y. Z., *Proteins* 2004, *55*, 66–76.

[30] Han, L. Y., Cai, C. Z., Lo, S. L., Chung, M. C., Chen, Y. Z., *RNA* 2004, *10*, 355–368.

[31] Bhasin, M., Raghava, G. P., *J. Biol. Chem.* 2004, *279*, 23262–23266.

[32] Cai, Y. D., Chou, K. C., *J. Proteome Res.* 2005, *4*, 967–971.

[33] Chou, K. C., *Bioinformatics* 2005, *21*, 10–19.

[34] Jensen, L. J., Gupta, R., Staerfeldt, H. H., Brunak, S., *Bioinformatics* 2003, *19*, 635–642.

[35] Bock, J. R., Gough, D. A., *Bioinformatics* 2003, *19*, 125–134.

[36] Martin, S., Roe, D., Faulon, J. L., *Bioinformatics* 2005, *21*, 218–226.

[37] Lo, S.L., Cai, C. Z., Chen, Y.Z. Chung, M. C. M., *Proteomics* 2005, *5*, 876–884.

[38] Sprinzak, E., Margalit, H., *J. Mol. Biol.* 2001, *311*, 681–692.

[39] Albert, I., Albert, R., *Bioinformatics* 2004, *20*, 3346–3352.

[40] Carr, A. M., Dorrington, S. M., Hindley, J., Phear, G. A. *et al.*, *Mol. Gen. Genet.* 1994, *245*, 628–635.

[41] de Lichtenberg, U., Jensen, T. S., Jensen, L. J., Brunak, S., *J. Mol. Biol.* 2003, *329*, 663–674.

[42] Ben-Hur, A., Noble, W. S., *Bioinformatics* 2005, 21 Suppl 1, i38–i46.

[43] King, R. D., Wise, P. H., Clare, A., *Bioinformatics* 2004, *20*, 1110–1118.

[44] King, R. D., Karwath, A., Clare, A., Dehaspe, L., *Yeast* 2000, *17*, 283–293.

[45] Cui, J., Han, L. Y., Cai, C. Z., Zheng, C. J. *et al.*, *J. Mol. Microbiol. Biotechnol.* 2005, *9*, 86–100.

[46] Han, L. Y., Cai, C. Z., Ji, Z. L., Chen, Y. Z., *Virology* 2005, *331*, 136–143.

[47] Bhasin, M., Raghava, G. P., *Nucleic Acids Res.* 2004, *32*, W383–389.

[48] Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H., Suwa, M., *Nucleic Acids Res.* 2005, *33*, W148–153.

[49] Cai, Y. D., Zhou, G. P., Chou, K. C., *Biophys. J.* 2003, *84*, 3257–3263.

[50] Wang, M., Yang, J., Liu, G. P., Xu, Z. J., Chou, K. C., *Protein Eng. Des. Sel.* 2004, *17*, 509–516.

[51] Bhardwaj, N., Langlois, R. E., Zhao, G., Lu, H., *Nucleic Acids Res.* 2005, *33*, 6486–6493.

[52] Lin, H. H., Han, L. Y., Zhang, H. L., Zheng, C. J. *et al.*, *J. Lipid Res.* 2006, *47*, 824–831.

[53] Dobson, P. D., Doig, A. J., *J. Mol. Biol.* 2005, *345*, 187–199.

[54] Lin, H. H., Han, L. Y., Cai, C. Z., Ji, Z. L., Chen, Y. Z., *Proteins* 2006, *62*, 218–231.

[55] Huang, N., Chen, H., Sun, Z., *Protein Eng. Des. Sel.* 2005, *18*, 365–368.

[56] Smialowski, P., Schmidt, T., Cox, J., Kirschner, A., Frishman, D., *Proteins* 2006, *62*, 343–355.

[57] Kumar, M., Verma, R., Raghava, G. P., *J. Biol. Chem.* 2006, *281*, 5357–5363.

[58] de Lichtenberg, U., Jensen, L. J., Fausboll, A., Jensen, T. S. *et al.*, *Bioinformatics* 2005, *21*, 1164–1171.

[59] Cai, Y. D., Doig, A. J., *Bioinformatics* 2004, *20*, 1292–1300.

[60] de Lichtenberg, U., Jensen, L. J., Brunak, S., Bork, P., *Science* 2005, *307*, 724–727.

[61] Zhang, Z., Kochhar, S., Grigorov, M. G., *Protein Sci.* 2005, *14*, 431–444.

[62] Hou, Y., Hsu, W., Lee, M. L., Bystroff, C., *Proteins* 2004, *57*, 518–530.

[63] Schneider, G., Wrede, P., *Biophys. J.* 1994, *66*, 335–344.

[64] Chou, K. C., *Biochem. Biophys. Res. Commun.* 2000, *278*, 477–483.

[65] Chou, K. C., Cai, Y. D., *Biochem. Biophys. Res. Commun.* 2004, *320*, 1236–1239.

[66] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. *et al.*, in Walker, J. M. (Ed.), *The Proteomics Protocols Handbook*, Humana Press, Totowa 2005, 571–607.

[67] Honeyman, M. C., Brusic, V., Stone, N. L., Harrison, L. C., *Nat. Biotechnol.* 1998, *16*, 966–969.

[68] Bhasin, M., Raghava, G. P., *Vaccine* 2004, *22*, 3195–3204.

[69] Tenzer, S., Peters, B., Bulik, S., Schoor, O. *et al.*, *Cell. Mol. Life Sci.* 2005, *62*, 1025–1037.

[70] Xue, Y., Yap, C. W., Sun, L. Z., Cao, Z. W. *et al.*, *J. Chem. Inf. Comput. Sci.* 2004, *44*, 1497–1505.

[71] Kawashima, S., Kanehisa, M., *Nucleic Acids Res.* 2000, *28*, 374.

[72] Cid, H., Bunster, M., Canales, M., Gazitua, F., *Protein Eng.* 1992, *5*, 373–375.

[73] Bhaskaran, R., Ponnuswammy, P. K., *Int. J. Pept. Protein Res.* 1988, *32*, 242–255.

[74] Charton, M., Charton, B. I., *J. Theor. Biol.* 1982, *99*, 629–644.

[75] Chothia, C., *J. Mol. Biol.* 1976, *105*, 1–12.

[76] Bigelow, C. C., *J. Theor. Biol.* 1967, *16*, 187–211.

[77] Charton, M., *J. Theor. Biol.* 1981, *91*, 115–123.

[78] Dayhoff, H., Calderone, H., *Altas of Protein Sequence and Structure*, Vol. 5. National Biomedical Research Foundation, Washington D.C. 1978, pp. 363–373.

[79] Feng, Z. P. Zhang, C. T., *J. Protein Chem.* 2000, *19*, 269–275.

[80] Moran, P. A., *Biometrika* 1950, *37*, 17–23.

[81] Sokal, R. R., Thomson, B. A., *Am. J. Phys. Anthropol.* 2006, *129*, 121–131.

[82] Grantham, R., *Science* 1974, *185*, 862–864.

[83] Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., Kim, S. H., *Proteins* 1999, *35*, 401–407.

[84] Dubchak, I., Muchnik, I., Holbrook, S. R., Kim, S. H., *Proc. Natl. Acad. Sci. USA* 1995, *92*, 8700–8704.

[85] Wang, D., Larder, B., *J. Infect. Dis.* 2003, *188*, 653–660.

[86] Draghici, S., Potter, R. B., *Bioinformatics* 2003, *19*, 98–107.

[87] Specht, D., *Neural Netw.* 1990, *3*, 109–118.

[88] Lipsitz, S. R., Parzen, M., *Biometrics* 1996, *52*, 291–298.

[89] Johnson, R., Wichern, D., *Applied multivariate statistical analysis*, Prentice Hall, Englewood Cliffs 1982.

[90] Fix, E., Hodges, J., *Discriminatory analysis: Non–parametric discrimination: Consistency Properties.*, USAF School of Aviation Medicine, Randolph Field 1951, pp. 261–279.

[91] Quinlan, J., *C4.5: programs for machine learning*, Morgan Kaufmann, San Mateo 1993.

[92] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, New York 1995.

[93] Cristianini, N., Shawe-Taylor, J., *An introduction to Support Vector Machines: and other kernel-based learning methods*, Cambridge University Press, New York 2000.

[94] Provost, F., Fawcett, T., Kohavi, R., *Proc. 15th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco 1998, pp. 445–453.

[95] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., Nielsen, H., *Bioinformatics* 2000, *16*, 412–424.

[96] Schomburg, I., Chang, A., Schomburg, D., *Nucleic Acids Res.* 2002, *30*, 47–49.

[97] Bull, H. B., Breese, K., *Arch. Biochem. Biophys.* 1974, *161*, 665–670.

[98] Lin, T. Y., Timasheff, S. N., *Protein Sci.* 1996, *5*, 372–381.

[99] Veropoulos, K., Campbell, C., Cristianini, N., in: Dean, T. (Ed.), *Proceedings of the International Joint Conference on Artificial Intelligence (UCAI99)*, Morgan Kaufmann, Stockholm 1999, pp. 55–60.

[100] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N. *et al.*, *Proc. Natl. Acad. Sci. USA* 2000, *97*, 262–267.

[101] Todeschini, R., Consonni, V., *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim 2002.

[102] Hall, L. H., Kellogg, G. E., Haney, D. N., *User's Guide*, eduSoft, LC, Ashland 2002.

[103] Xue, Y., Li, Z. R., Yap, C. W., Sun, L. Z. *et al.*, *J. Chem. Inf. Comput. Sci.* 2004, *44*, 1630–1638.

[104] Al–Shahib, A., Breitling, R., Gilbert, D., *Int. J. Neural Syst.* 2005, *15*, 259–275.

[105] Al–Shahib, A., Breitling, R., Gilbert, D., *Appl. Bioinformatics* 2005, *4*, 195–203.

[106] Furlanello, C., Serafini, M., Merler, S., Jurman, G., *Neural Netw.* 2003, *16*, 641–648.

[107] Li, H., Ung, C., Yap, C., Xue, Y. *et al.*, *Chem. Res. Toxicol.* 2005, *18*, 1071–1080.

[108] Yap, C. W., Chen, Y. Z., *J. Chem. Inf. Model.* 2005, *45*, 982–992.

[109] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G. *et al.*, *Nucleic Acids Res.* 2000, *28*, 235–242.

[110] Bateman, A., Coin, L., Durbin, R., Finn, R. D. *et al.*, *Nucleic Acids Res.* 2004, *32*, D138–141.

[111] Lee, T. Y., Huang, H. D., Hung, J. H., Huang, H. Y. *et al.*, *Nucleic Acids Res.* 2006, *34*, D622–627.

[112] Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E. *et al.*, *Nucleic Acids Res.* 2003, *31*, 28–33.

[113] Dorazilova, V., Vedralova, J., *Cesk. Patol.* 1992, *28*, 245–247.

[114] Barker, W. C., Garavelli, J. S., McGarvey, P. B., Marzec, C. R. *et al.*, *Nucleic Acids Res.* 1999, *27*, 39–43.

[115] Benson, D. A., Karsch–Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., *Nucleic Acids Res.* 2004, *32*, D23–26.

[116] Chalmel, F., Lardenois, A., Thompson, J. D., Muller, J. *et al.*, *Bioinformatics* 2005, *21*, 2095–2096.

[117] Kanehisa, M., Goto, S., Hattori, M., Aoki–Kinoshita, K. F. *et al.*, *Nucleic Acids Res.* 2006, *34*, D354–357.

[118] Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. *et al.*, *Nucleic Acids Res.* 2004, *32*, D226–229.

[119] Holm, L., Sander, C., *Science* 1996, *273*, 595–603.

[120] Attwood, T. K., Croning, M. D., Flower, D. R., Lewis, A. P. *et al.*, *Nucleic Acids Res.* 2000, *28*, 225–227.

[121] Bairoch, A., *Nucleic Acids Res.* 2000, *28*, 304–305.

[122] Schomburg, I., Chang, A., Ebeling, C., Gremse, M. *et al.*, *Nucleic Acids Res.* 2004, *32*, D431–433.

[123] Saier, M. H., Jr., Tran, C. V., Barabote, R. D., *Nucleic Acids Res.* 2006, *34*, D181–186.

[124] Ren, Q., Kang, K. H., Paulsen, I. T., *Nucleic Acids Res.* 2004, *32*, D284–288.

[125] Yan, Q., Sadee, W., *AAPS PharmSci.* 2000, *2*, E20.

[126] Horn, F., Bettler, E., Oliveira, L., Campagne, F. *et al.*, *Nucleic Acids Res.* 2003, *31*, 294–297.

[127] Chen, X., Ji, Z. L., Chen, Y. Z., *Nucleic Acids Res.* 2002, *30*, 412–415.

[128] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P. *et al.*, *Nucleic Acids Res.* 2002, *30*, 303–305.

[129] Ji, Z. L., Chen, X., Zhen, C. J., Yao, L. X. *et al.*, *Nucleic Acids Res.* 2003, *31*, 255–257.

[130] Roth, B. L., Kroeze, W. K., Patel, S., Lopez, E., *Neuroscientist* 2000, *6*, 252–262.

[131] Zhang, J., Aizawa, M., Amari, S., Iwasawa, Y. *et al.*, *Comput. Biol. Chem.* 2004, *28*, 401–407.

[132] Chen, X., Ji, Z. L., Zhi, D. G., Chen, Y. Z., *Comput. Chem.* 2002, *26*, 661–666.

