Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity

H. H. Lin,* L. Y. Han,* H. L. Zhang,* C. J. Zheng,* B. Xie,* and Y. Z. Chen,[1,*,†]

Bioinformatics and Drug Design Group,* Department of Computational Science, National University of Singapore, Singapore 117543; and Shanghai Center for Bioinformatics Technology,† Shanghai 200235, People's Republic of China

Abstract Lipid binding proteins play important roles in signaling, regulation, membrane trafficking, immune response, lipid metabolism, and transport. Because of their functional and sequence diversity, it is desirable to explore additional methods for predicting lipid binding proteins irrespective of sequence similarity. This work explores the use of support vector machines (SVMs) as such a method. SVM prediction systems are developed using 14,776 lipid binding and 133,441 nonlipid binding proteins and are evaluated by an independent set of 6,768 lipid binding and 64,761 nonlipid binding proteins. The computed prediction accuracy is 78.9, 79.5, 82.2, 79.5, 84.4, 76.6, 90.6, 79.0, and 89.9% for lipid degradation, lipid metabolism, lipid synthesis, lipid transport, lipid binding, lipopolysaccharide biosynthesis, lipoprotein, lipoyl, and all lipid binding proteins, respectively. The accuracy for the nonmember proteins of each class is 99.9, 99.2, 99.6, 99.8, 99.9, 99.8, 98.5, 99.9, and 97.0%, respectively. Comparable accuracies are obtained when homologous proteins are considered as one, or by using a different SVM kernel function. Our method predicts 86.8% of the 76 lipid binding proteins nonhomologous to any protein in the Swiss-Prot database and 89.0% of the 73 known lipid binding domains as lipid binding. These findings suggest the usefulness of SVMs for facilitating the prediction of lipid binding proteins. Our software can be accessed at the SVMProt server (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi).—Lin, H. H., L. Y. Han, H. L. Zhang, C. J. Zheng, B. Xie, and Y. Z. Chen. **Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. *J. Lipid Res.* 2006.** 47: **824–831.**

Supplementary key words lipid-protein interactions • lipid-modifying enzymes • lipid metabolism • support vector machine

Lipid binding proteins play important roles in cell signaling and membrane trafficking (1), lipid metabolism and transport (2, 3), innate immune responses to bacterial infections (4), and the regulation of gene expression and cell growth (5). Prediction of the functional roles of lipid binding proteins is important for facilitating the study of various biological processes and the search for new therapeutic targets. Intensive efforts have been directed at the study of the genetics of lipid binding (3, 5) and the molecular mechanism of lipid-protein interactions, which provide useful clues about sequence features, structural characteristics, domains, physicochemical properties, and kinetic data related to lipid binding and metabolism (6–13), which can be explored for developing methods to predict the function of lipid binding proteins.

At present, prediction of the function of lipid binding proteins is primarily based on sequence similarity and clustering methods (14) and the identification of sequence signals and motifs (15–19). It is known that many genomes contain substantial percentages of the putative protein-coding open reading frames, which are nonhomologous to any protein of known function (20, 21). Therefore, it is desirable to explore additional methods that predict protein function irrespective of sequence similarity. A statistical learning method, the use of support vector machines (SVMs), has been used successfully to predict the functional classes of molecule binding proteins such as RNA binding proteins (22, 23), DNA binding proteins (23), and transporters (24) irrespective of sequence similarity from sequence-derived structural and physicochemical properties. SVMs also showed a certain level of capability for predicting novel proteins that have no known similarity to any other proteins (25, 26). It is thus of interest to explore SVMs to predict the functional classes of lipid binding proteins.

Lipid binding proteins are diverse in sequence, structure, and function (6–13). Nontheless, lipid recognition by proteins is primarily mediated by some combination of a number of structural and physicochemical features, including

Manuscript received 6 December 2005 and in revised form 17 January 2006.

Published, JLR Papers in Press, January 27, 2006.
DOI 10.1194/jlr.M500530-JLR200

[1] To whom correspondence should be addressed.
e-mail: yzchen@cz3.nus.edu.sg

conserved fold elements (5), specific lipid binding site architectures (6) and recognition motifs (7, 13), ordered hydrophobic and polar contacts between lipid and protein (8), and multiple noncovalent interactions from protein residues to lipid head groups and hydrophobic tails (13). To some extent, these lipid-protein binding features are similar to those of other molecule binding features of proteins, such as RNA binding proteins, DNA binding proteins, and transporters. For instance, RNA binding proteins are also diverse in sequence, structure, and function, and their binding capabilities are mediated by certain classes of RNA binding domains and motifs (27–30). Therefore, it is expected that SVMs are also applicable to the prediction of the functional classes of lipid binding proteins.

Here, we explore the use of SVMs for developing prediction systems for eight lipid binding classes and for all lipid binding proteins. These classes 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, such as farnesyl, palmitate, and myristate), and lipoyl (proteins containing at least one lipoyl binding domain). In addition to the estimate of prediction accuracy using an independent set of proteins, the performance of our developed SVM prediction systems is further evaluated by four additional tests to determine the usefulness of SVMs to predict novel lipid binding proteins and the applicability of other kernel functions. One is the evaluation of the prediction accuracies when homologous proteins are considered as one. The second is the prediction of lipid binding proteins nonhomologous to any protein in the Swiss-Prot database (31). The third is to study whether the known lipid binding domains can be predicted as lipid binding by our SVM systems. The fourth is to study the performance of SVMs with a different kernel function.

METHODS

Selection of lipid binding and nonlipid binding proteins

All lipid binding proteins used in this study are from a comprehensive search of the Swiss-Prot database at http://www.expasy.uniprot.org (31). A total of 10,815 lipid binding protein sequences are obtained. The distribution of most of these proteins in specific lipid binding classes is 873, 659, 2,383, 341, 607, 565, 5,097, and 204 in the lipid degradation, lipid metabolism, lipid synthesis, lipid transport, lipid binding, lipopolysaccharide biosynthesis, lipoprotein, and lipoyl classes, respectively. Some proteins are found to belong to more than one class. The distribution of all these proteins in different kingdoms and in the top 10 host species is given in **Table 1**, and that of some classes of lipid binding proteins is given in **Table 2**. From these two tables, one finds that these proteins are from a diverse range of species and that all species appear to be fairly adequately represented.

It is likely that not all of the identified lipid binding protein sequences that belong to each of these eight lipid binding classes are explicitly specified in the protein sequence database. Effort is made to manually check all of the selected lipid binding protein sequences to determine whether or not some of them belong to a specific class. It is expected that some of these proteins may still be missed and thus are not included in their respective classes.

All distinct members in each class are used to construct a positive data set for the corresponding SVM classification system. A negative data set, representing nonclass members, is selected by a well-established procedure (26, 32, 33), such that all proteins are grouped into domain families (34) and the representative proteins of those families unrelated to the specific lipid binding class are used as negative samples. Members in the other lipid binding classes are included in the negative data set if they are unrelated to the class being studied. These data sets are divided into separate training, testing, and independent evaluation sets in such a way that all of the distinct proteins, the remaining distinct proteins, and the rest are distributed in the training, testing, and independent evaluation sets, respectively. Statistical data for the members and nonmembers in each data set of each lipid binding class are given in **Table 3**.

TABLE 1. Distribution of lipid binding proteins in different kingdoms and in the top 10 host species of each kingdom

Variable	Kingdom			
	Viridae	Eukaryota	Bacteria	Archaea
Number of proteins in kingdom	837	5,560	4,183	235
Top 10 species and number of proteins in each species	*Autographa californica* nuclear polyhedrosis virus (12)	*Homo sapiens* (758)	*Escherichia coli* (254)	*Methanococcus jannaschii* (73)
	Variola virus (6)	*Mus musculus* (622)	*Haemophilus influenzae* (117)	*Archaeoglobus fulgidus* (32)
	Vaccinia virus (strain Copenhagen) (6)	*Rattus norvegicus* (373)	*Salmonella typhimurium* (106)	*Pyrococcus horikoshii* (14)
	Vaccinia virus (strain Western Reserve/WR) (6)	*Arabidopsis thaliana* (197)	*Bacillus subtilis* (100)	*Aeropyrum pernix* (11)
	Orgyia pseudotsugata multicapsid polyhedrosis virus (4)	*Bos taurus* (189)	*Mycobacterium bovis* (77)	*Pyrococcus abyssi* (11)
	Reovirus type 3 (strain Dearing) (4)	*Saccharomyces cerevisiae* (186)	*Mycobacteriumtuberculosis* (74)	*Sulfolobus solfataricus* (9)
	Vaccinia virus (strain Ankara) (4)	*Gallus gallus* (105)	*Escherichia coli* O157:H7 (70)	*Pyrococcus furiosus* (8)
	Reovirus type 2 (strain D5/Jones) (4)	*Caenorhabditis elegans* (100)	*Mycoplasma pneumoniae* (70)	*Methanobacterium thermoautotrophicum* (8)
	Human immunodeficiency virus type 2 (isolate CAM2) (3)	*Sus scrofa* (93)	*Shigella flexneri* (63)	*Methanosarcina mazei* (8)
	Human immunodeficiency virus type 1 (isolate PV22) (3)	*Canis familiaris* (89)	*Vibrio cholerae* (54)	*Thermoplasma acidophilum* (7)

TABLE 2. Distribution of lipid binding proteins involved in lipid transport, lipid synthesis, and lipid degradation in different kingdoms and in the top 10 host species

Variable	Lipid Transport		Lipid Synthesis		Lipid Degradation	
	Kingdom or Species	No. of Proteins	Kingdom or Species	No. of Proteins	Kingdom or Species	No. of Proteins
Protein distribution in kingdom	Archaea	—	Archaea	63	Archaea	10
	Bacteria	53	Bacteria	1,447	Bacteria	301
	Eukaryota	288	Eukaryota	842	Eukaryota	562
	Viridae	—	Viridae	31	Viridae	—
Protein distribution in top 10 species	*Homo sapiens*	50	*Homo sapiens*	81	*Homo sapiens*	47
	Mus musculus	34	*Mus musculus*	71	*Mus musculus*	40
	Rattus norvegicus	18	*Arabidopsis thaliana*	66	*Rattus norvegicus*	34
	Bos taurus	12	*Rattus norvegicus*	52	*Austrelaps superbus*	19
	Sus scrofa	10	*Saccharomyces cerevisiae*	42	*Bos taurus*	17
	Oryctolagus cuniculus	9	*Escherichia coli*	29	*Candida albicans*	15
	Gallus gallus	8	*Schizosaccharomyces pombe*	28	*Saccharomyces cerevisiae*	15
	Macaca fascicularis	8	*Oryza sativa*	25	*Arabidopsis thaliana*	14
	Saccharomyces cerevisiae	8	*Salmonella typhimurium*	21	*Laticauda semifasciata*	12
	Canis familiaris	6	*Haemophilus influenzae*	20	*Bungarus multicinctus*	12

Derivation of structural and physicochemical properties from protein sequence

Construction of the feature vector for each protein is based on the formulae used in the prediction of RNA binding proteins (33), protein-protein interaction (35), protein fold recognition (36), and protein functional family prediction (32). Given the sequence of a protein, its amino acid composition and the properties of every constituent amino acid are computed and then used to generate this vector. The computed amino acid properties include hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility (32).

For each of these properties, amino acids are divided into three groups such that those in a particular group are regarded to have approximately the same property. For instance, amino acids can be divided into hydrophobic (CVLIMFW), neutral (GASTPHY), and polar (RKEDQN) groups. Three descriptors, composition (C), transition (T), and distribution (D), are introduced to describe the global composition of each of these properties. 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. T characterizes the percentage frequency with which amino acids of a particular property are followed by amino acids of a different property. D measures the chain length within which the first, 25, 50, 75, and 100% of the amino acids of a particular property are located.

A hypothetical protein sequence, AEAAAEAEEAAAAAEAEEEAAEEAEEEAAE, as shown in **Fig. 1**, has 16 alanines (n1 = 16) and 14 glutamic acids (n2 = 14). The composition for these two amino acids is $n1 \times 100.00/(n1 + n2) = 53.33$ and $n2 \times 100.00/(n1 + n2) = 46.67$, respectively. There are 15 transitions from A to E or from E to A in this sequence, and the percentage frequency of these transitions is $(15/29) \times 100.00 = 51.72$. The first, 25, 50, 75, and 100% of As are located within the first 1, 5, 12, 20, and 29 residues, respectively. The D descriptor for As is thus $1/30 \times 100.00 = 3.33$, $5/30 \times 100.00 = 16.67$, $12/30 \times 100.00 = 40.0$, $20/30 \times 100.00 = 66.67$, and $29/30 \times 100.00 = 96.67$. Likewise, the D descriptor for Es is 6.67, 26.67, 60.0, 76.67, and 100.0. Overall, the amino acid composition descriptors for this sequence are C = 53.33 and 46.67, T = 51.72, and D = 3.33, 16.67, 40.0, 66.67, 96.67, 6.67, 26.67, 60.0, 76.67, and 100.0. Descriptors for other properties can be computed by a similar procedure.

Overall, there are 21 elements representing these three descriptors: 3 for C, 3 for T, and 15 for D. The feature vector of a protein is constructed by combining the 21 elements of all of these properties and the 20 elements of amino acid composition in sequential order.

TABLE 3. Statistics of the training, testing, and independent evaluation set, and prediction accuracy of individual classes of lipid binding proteins and all lipid binding proteins

Lipid Binding Protein Class	Training Set		Testing Set				Independent Evaluation Set						
			Positive		Negative		Positive			Negative			
	Positive	Negative	TP	FN	TN	FP	TP	FN	SE	TN	FP	SP	Q
Lipid degradation	403	1,775	232	1	13,629	6	187	50	78.9%	7,696	5	99.9%	99.3%
Lipid metabolism	293	1,969	155	50	13,276	124	128	33	79.5%	7,632	64	99.2%	98.8%
Lipid synthesis	891	2,607	722	67	12,615	55	578	125	82.2%	7,605	30	99.6%	98.1%
Lipid transport	153	2,109	87	23	13,436	26	62	16	79.5%	7,712	15	99.8%	99.6%
Lipid binding	274	1,530	165	1	13,918	8	141	26	84.4%	7,714	10	99.9%	99.5%
Lipopolysaccharide biosynthesis	285	10,837	134	9	4,590	0	105	32	76.6%	7,710	13	99.8%	99.6%
Lipoprotein	1,648	5,065	1,724	68	8,634	199	1,501	156	90.6%	5,459	81	98.5%	96.7%
Lipoyl	72	1,987	58	12	13,529	17	49	13	79.0%	7,758	11	99.9%	99.7%
All lipid binding proteins	3,232	3,701	3,966	51	7,763	36	3,205	361	89.9%	5,086	160	97.0%	94.1%

The predicted results are given in true positive (TP), false negative (FN), true negative (TN), false positive (FP), sensitivity (SE) [=TP/(TP + FN)] (accuracy for class members), specificity (SP) [=TN/(TN + FP)] (accuracy for nonmembers), and overall accuracy (Q) [=(TN + TP)/(TP + FN + TN + FP)]. The number of members and nonmembers in the testing and independent evaluation sets is given by TP + FN or TN + FP, respectively.

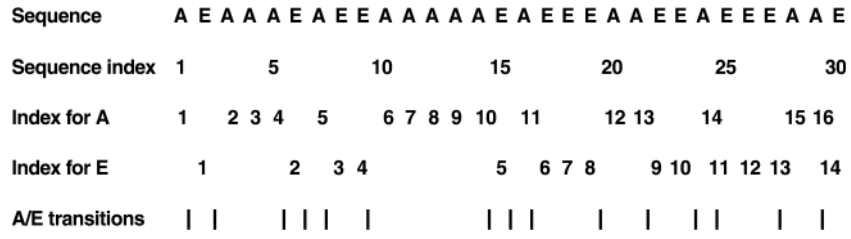

Fig. 1. Sequence of a hypothetic protein for illustration of the derivation of the feature vector of a protein. The sequence index indicates the position of an amino acid in the sequence. The index for each type of amino acids in the sequence (A or E) indicates the position of the first, second, third … of that type of amino acid (the position of the first, second, third… A is at 1, 3, 4…). A/E transition indicates the positions of AE or EA pairs in the sequence.

There is some level of overlap in the descriptors for hydrophobicity, polarity, and surface tension. Thus, the dimensionality of the feature vectors may be reduced by principal component analysis. Our own study suggests that the use of principal component analysis-reduced feature vectors only moderately improves the accuracy for some of the families. Thus, it is unclear to what extent this overlap affects the accuracy of SVM classification. It is noted that reasonably accurate results have been obtained using these overlapping descriptors in various protein classification studies (32, 35–38).

SVM method

The algorithms of SVM and its applications to proteins are extensively described in the literature (32, 33, 39). Thus, only a brief description is given here. A linear SVM constructs a hyperplane that separates two different classes of feature vectors with a maximum margin. One class represents lipid binding proteins, and the other represents nonlipid binding proteins. This hyperplane is constructed by finding a vector $\mathbf{w}$ and a parameter b that minimizes $\|\mathbf{w}\|^2$, which satisfies the following conditions: $\mathbf{w} \times \mathbf{x_i} + b \geq +1$, for $y_i = +1$ (positive class), and $\mathbf{w} \times \mathbf{x_i} + b \leq -1$, for $y_i = -1$ (negative class). Here, $\mathbf{x_i}$ is a feature vector, y_i is the group index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b| / \|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin, and $\|\mathbf{w}\|^2$ is the Euclidean norm of $\mathbf{w}$.

A nonlinear SVM projects feature vectors into a high-dimensional feature space using a kernel function such as the Gaussian kernel function $K(\mathbf{x_i}, \mathbf{x_j}) = e^{-|\mathbf{xj} - \mathbf{xi}|2/2\sigma 2}$. 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 $sign[(\mathbf{w} \times \mathbf{x}) + b]$; a positive or negative value indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

The performance of SVM has been measured by the positive, negative, and overall prediction accuracies $P_p = TP/(TP + FN)$, $P_n = TN/(TN + FP)$, and $P = (TP + TN)/N$, which correspond to the accuracies for proteins of a lipid binding class, nonmembers of the class, and all members and nonmembers of the class, respectively. Here, TP, TN, FP, and FN are the number of true positives (true member), true negatives (true nonmember), false positives (false member), and false negatives (false nonmember), respectively, and N is the total number of proteins studied.

RESULTS AND DISCUSSION

Overall prediction accuracy

The statistics of the data sets and prediction results for specific lipid binding classes and all lipid binding proteins are given in Table 3. In this table, TP, FN, TN, FP, SE, and SP stand for true positive (correctly predicted lipid binding proteins of a specific class), false negative (specific class of lipid binding proteins incorrectly predicted as nonclass members), true negative (correctly predicted nonclass members), false positive (nonclass members incorrectly predicted as a specific class of lipid binding proteins), predicted sensitivity (accuracy for members in each lipid binding class), and predicted specificity (accuracy for nonmembers of each lipid binding class), respectively. The SEs for the lipid degradation, lipid metabolism, lipid synthesis, lipid transport, lipid binding, lipopolysaccharide biosynthesis, lipoprotein, lipoyl, and all lipid binding proteins are 78.9, 79.5, 82.2, 79.5, 84.4, 76.6, 90.6, 79.0, and 89.9%, respectively. The corresponding SPs are 99.9, 99.2, 99.6, 99.8, 99.9, 99.8, 98.5, 99.9, and 97.0%, respectively. When homologous proteins are considered as one, the SEs become 76.9, 77.9, 80.9, 79.7, 83.1, 74.2, 90.4, 78.6, and 89.8% and the SPs become 99.9, 99.1, 99.6, 99.8, 99.9, 99.8, 98.6, 99.9, and 96.9%, respectively. Overall, the SEs are reduced slightly and the SPs are almost unchanged compared with the results derived from the use of all proteins.

A direct comparison with results from previous lipid binding protein prediction studies may not be most appropriate because of the differences in the protein classes predicted, data sets, protein descriptors, prediction methods, and parameters. Nonetheless, a tentative comparison may provide some crude estimate regarding the level of accuracy of our method with respect to those achieved by other studies of lipid binding proteins. The reported SEs and SPs of other studies are in the range of 92∼97% and ∼99% for the lipoprotein proteins (17, 18) and 80∼95% and 99.2∼99.9% for lipid modification proteins (16). Although our results are comparable to those of other studies, a significantly higher number, and thus more diverse range, of proteins is covered in our studies.

The prediction accuracy of the nonmembers of each lipid binding class appears to be better than that of the members. The higher prediction accuracy for nonmembers likely results from the availability of a more diverse set of nonmembers than that of members, which enables the SVM to perform a better statistical learning for recognition of nonmembers. Based on the statistics provided on the

Pfam database webpage (34), there are >7,000 families of proteins, from which one can generate a diverse set of nonmembers for each DNA binding class.

Because of differences in the numbers of members and nonmembers in each class, there is an imbalance between each data set. SVMs based on imbalanced data sets tend to produce feature vectors that push the hyperplane toward the side with the smaller number of data (40), which can lead to reduced accuracy for the set with either a smaller number of samples or less diversity. This might partly explain why the prediction accuracy for members is generally lower than that for nonmembers. However, it is inappropriate to simply reduce the number of nonmembers to artificially match that of members, because this compromises the diversity needed to fully represent all nonmembers. Computational methods for readjusting the biased shift of the hyperplane are being explored (41). Application of these methods may help to improve SVM prediction accuracy in this and other cases involving unbalanced data.

Prediction of novel lipid binding proteins

One particular application of our SVM classification systems is for the prediction of novel lipid binding proteins that are nonhomologous to other proteins. To test this capability, the Swiss-Prot database (31) is searched for lipid binding proteins having no single homologous protein in the database based on PSI-BLAST (14) results. A similarity E-value threshold of 0.1 is used for the homolog search to ensure the maximum exclusion of proteins that have a homolog. Those proteins found in the SVM training sets are then removed. As shown in **Table 4**, 76 proteins are found by this process, and 66 or 86.8% of these proteins are correctly predicted as lipid binding by our SVM classification systems. Therefore, our SVM classification systems appear to show reasonably good capability for predicting novel lipid binding proteins based on the set of proteins tested.

Prediction of proteins with specific structural characteristics

A number of lipid binding proteins contain lipid binding domains or motifs (7). Several families of such lipid binding proteins have been documented, and examples of these families are TIM, PP binding, and GCV_H. These families have distinctive structural features responsible for lipid recognition and binding. Thus, the performance of SVM classification of lipid binding proteins can be evaluated by examining whether or not proteins containing one of these domains or motifs can be correctly classified as lipid binding proteins.

A search of protein family and sequence databases shows that there are 227, 184, and 139 lipid binding protein sequences known to contain the TIM, PP binding, and GCV_H domains, respectively. The majority of these sequences are included in the training and testing set of all DNA binding proteins. In the corresponding independent evaluation set, there are 81, 27, and 30 sequences containing the TIM, PP binding, and GCV_H domains, respectively. Most of these protein sequences are correctly classified as lipid binding by SVMs. There are only one, one, and two misclassified sequences in the TIM, PP binding, and GCV_H domain families, respectively. Thus, our results show the capability of our SVM prediction systems for recognizing these lipid binding proteins. The incorrectly predicted protein sequences are triosephosphate isomerase (fragment), putative acyl carrier protein, mitochondrial precursor, glycine cleavage system H protein, mitochondrial precursor (fragment), and probable glycine cleavage system H protein 2, mitochondrial precursor. Most of these incorrectly predicted sequences are fragments. Therefore, sequence incompleteness appears

TABLE 4. Prediction results of novel lipid binding proteins by SVMProt, where + represents proteins correctly predicted as lipid binding proteins and − represents proteins incorrectly predicted as nonlipid binding proteins

Swiss-Prot AC	Prediction Status	Swiss-Prot AC	Prediction Status	Swiss-Prot AC	Prediction Status	Swiss-Prot AC	Prediction Status
O13547	+	P16055	+	P39907	+	P77339	+
O15255	+	P18149	+	P39910	+	P77717	+
O32528	+	P18164	+	P41052	+	P83408	−
O59715	+	P18952	−	P41069	−	P97029	+
O66867	+	P19411	+	P41365	+	Q01821	+
O67301	+	P19412	+	P42461	+	Q03490	+
O67672	+	P19478	+	P42708	+	Q05903	+
O83276	−	P19833	+	P43497	+	Q08906	+
O83469	+	P25666	+	P46122	+	Q46122	+
O83516	+	P26471	−	P54660	+	Q46670	+
O83691	−	P27126	+	P55428	+	Q46835	+
O83811	+	P27832	+	P55703	+	Q47499	+
P07096	+	P29723	+	P65302	+	Q50675	+
P08452	+	P32323	+	P65310	+	Q53728	−
P08472	+	P33219	+	P65316	−	Q54313	+
P0A0V1	+	P37056	−	P70837	−	Q56032	+
P0A1X3	+	P37261	+	P75734	+	Q94BT2	+
P11910	+	P37748	+	P75737	+	Q9CJU4	+
P12729	+	P38371	+	P75818	+	Q9CLP1	+

AC, accession number.

to be a factor that partially contributes to the incorrect prediction of these sequences by SVMs.

Prediction performance for lipid binding domains

Some lipid binding proteins are known to contain multiple domains that include a lipid binding domain plus one or more domains characterized by DNA binding, protein-protein interaction, and other motifs (42–45). Our SVM prediction systems are trained using physicochemical properties derived from the entire protein sequence. There is a need to evaluate how the inclusion of all of these other "extra" domains may affect the prediction performance of our SVM systems. For such a purpose, our SVM systems are tested to determine to what extent they can predict known lipid binding domains as lipid binding without having to include representatives of these domains in our training sets. Lipid binding domains are searched from the Pfam database (34) using key word "lipid" against the Pfam, Prosite, and UniProt databases, followed by manual evaluation of the hits to select those with such annotations as involvement in lipid synthesizing, transporting, metabolizing, transferring, and degrading, interaction with lipid, and lipoprotein. A total of 73 distinct lipid binding domains are selected from this process, which include 23 domains in multidomain lipid binding proteins. We found that 89.0% and 82.6% of these are predicted as lipid binding. Moreover, 87.2% of the 632 multidomain lipid binding proteins in our independent set are correctly predicted. Hence, the inclusion of extra domains appears to have a limited effect on the performance of our developed SVM systems, which show a certain level of capability to predict lipid binding domains as well as lipid binding proteins.

SVM prediction performance using a different kernel function

Apart from the Gaussian kernel function of sequence-derived physicochemical properties used in this work, several other kernel functions have been developed and applied for SVM analysis of proteins and DNAs (46–54). It is of interest to test the usefulness of some of these kernel functions for predicting lipid binding proteins. The string-kernel function has been used extensively and has shown promising potential for protein and DNA studies (46, 47). 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. In this work, this kernel function is used to develop three SVM systems to predict the lipid degradation, lipid metabolism, and lipid synthesis protein classes. Spectrum kernel with mismatches (53) is used to generate the string-kernel for each protein. Testing results using the independent set of proteins for each class show that the SEs are 77.2, 75.8, and 77.8% and the SPs are 97.6, 96.4, and 94.2% for each of these classes, respectively. Thus, comparable prediction performance can be achieved using string-kernel SVMs, which suggests the usefulness of this and other kernel functions for SVM prediction of lipid binding proteins.

Contribution of feature properties to the classification of lipid binding proteins

In this work, nine feature properties are used to describe physicochemical characteristics of each protein, which have been used routinely for the prediction of RNA binding proteins (55) and other proteins (32, 35–38). It has been reported that not all feature vectors contribute equally to the classification of proteins; some have been found to play relatively more prominent roles than others in specific aspects of proteins (36). Therefore, it is of interest to examine which feature properties play more prominent roles in the classification of lipid binding proteins.

In an earlier study, the contributions of individual feature properties to protein classification were investigated by separately conducting classification using each feature property (36). The same method was used here. An analysis of the classification of the group of all lipid binding proteins suggests that, in order of prominence, polarity, hydrophobicity, amino acid composition, and solvent accessibility play more prominent roles than other feature properties. 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 (56). High-affinity lipid binding sites in some proteins appear to be located at sequence segments with specific amino acid composition (57), and specific sequence motifs have been used to predict lipid binding proteins (15–19). A study of apolipophorin III in lipid-free and phospholipid-bound states showed that lipid binding involves increased solvent accessibility, as a result of gross tertiary structural reorganization (58). Therefore, our prediction results are consistent with these experimental findings.

Conclusion

SVMs appear to be potentially useful tools for the prediction of lipid binding proteins of different classes. The prediction accuracy may be further enhanced with the expansion of our knowledge about lipid binding proteins, particularly for those small lipid binding classes, more refined representation of the structural and physicochemical properties of proteins, and the improvement of prediction algorithms, such as better treatment of an imbalanced data set. The SVM-derived lipid binding protein classification systems developed in this work can be accessed, free of charge for academic use, at the SVMProt server http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

This work was supported in part by grants from Singapore Academic Fund R-151-000-034-112, R-151-000-036-112, Shanghai Commission for Science and Technology (04DZ19850), and the "973" National Key Basic Research Program of China (2004CB720103).

REFERENCES

1. Downes, C. P., A. Gray, and J. M. Lucocq. 2005. Probing phosphoinositide functions in signaling and membrane trafficking. *Trends Cell Biol.* **15:** 259–268.
2. Glatz, J. F., J. J. Luiken, M. van Bilsen, and G. J. van der Vusse. 2002. Cellular lipid binding proteins as facilitators and regulators of lipid metabolism. *Mol. Cell. Biochem.* **239:** 3–7.
3. Haunerland, N. H., and F. Spener. 2004. Fatty acid-binding proteins—insights from genetic manipulations. *Prog. Lipid Res.* **43:** 328–349.
4. Bingle, C. D., and C. J. Craven. 2004. Meet the relatives: a family of BPI- and LBP-related proteins. *Trends Immunol.* **25:** 53–55.
5. Bernlohr, D. A., M. A. Simpson, A. V. Hertzel, and L. J. Banaszak. 1997. Intracellular lipid binding proteins and their genes. *Annu. Rev. Nutr.* **17:** 277–303.
6. Niggli, V. 2001. Structural properties of lipid binding sites in cytoskeletal proteins. *Trends Biochem. Sci.* **26:** 604–611.
7. Balla, T. 2005. Inositol-lipid binding motifs: signal integrators through protein-lipid and protein-protein interactions. *J. Cell Sci.* **118:** 2093–2104.
8. Pebay-Peyroula, E., and J. P. Rosenbusch. 2001. High-resolution structures and dynamics of membrane protein-lipid complexes: a critique. *Curr. Opin. Struct. Biol.* **11:** 427–432.
9. Fyfe, P. K., A. V. Hughes, P. Heathcote, and M. R. Jones. 2005. Proteins, chlorophylls and lipids: X-ray analysis of a three-way relationship. *Trends Plant Sci.* **10:** 275–282.
10. Bolanos-Garcia, V. M., and R. N. Miguel. 2003. On the structure and function of apolipoproteins: more than a family of lipid binding proteins. *Prog. Biophys. Mol. Biol.* **83:** 47–68.
11. Hanhoff, T., C. Lucke, and F. Spener. 2002. Insights into binding of fatty acids by fatty acid binding proteins. *Mol. Cell. Biochem.* **239:** 45–54.
12. Weisiger, R. A. 2002. Cytosolic fatty acid binding proteins catalyze two distinct steps in intracellular transport of their ligands. *Mol. Cell. Biochem.* **239:** 35–43.
13. Palsdottir, H., and C. Hunte. 2004. Lipids in membrane protein structures. *Biochim. Biophys. Acta.* **1666:** 2–18.
14. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.
15. Gonnet, P., and F. Lisacek. 2002. Probabilistic alignment of motifs with sequences. *Bioinformatics.* **18:** 1091–1101.
16. Eisenhaber, F., B. Eisenhaber, W. Kubina, S. Maurer-Stroh, G. Neuberger, G. Schneider, and M. Wildpaner. 2003. Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Pi, NMT and PTS1. *Nucleic Acids Res.* **31:** 3631–3634.
17. Juncker, A. S., H. Willenbrock, G. Von Heijne, S. Brunak, H. Nielsen, and A. Krogh. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12:** 1652–1662.
18. Gonnet, P., K. E. Rudd, and F. Lisacek. 2004. Fine-tuning the prediction of sequences cleaved by signal peptidase II: a curated set of proven and predicted lipoproteins of *Escherichia coli* K-12. *Proteomics.* **4:** 1597–1613.
19. Eisenhaber, B., F. Eisenhaber, S. Maurer-Stroh, and G. Neuberger. 2004. Prediction of sequence signals for lipid post-translational modifications: insights from case studies. *Proteomics.* **4:** 1614–1625.
20. Kalinowski, J., B. Bathe, D. Bartels, N. Bischoff, M. Bott, A. Burkovski, N. Dusch, L. Eggeling, B. J. Eikmanns, L. Gaigalat, et al. 2003. The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J. Biotechnol.* **104:** 5–25.
21. Henne, A., H. Bruggemann, C. Raasch, A. Wiezer, T. Hartsch, H. Liesegang, A. Johann, T. Lienard, O. Gohl, R. Martinez-Arias, et al. 2004. The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nat. Biotechnol.* **22:** 547–553.
22. Birch, P. J., L. V. Dekker, I. F. James, A. Southan, and D. Cronk. 2004. Strategies to identify ion channel modulators: current and novel approaches to target neuropathic pain. *Drug Discov. Today.* **9:** 410–418.
23. Cai, Y. D., and S. L. Lin. 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta.* **1648:** 127–133.
24. Lin, H. H., L. Y. Han, C. Z. Cai, Z. L. Ji, and Y. Z. Chen. 2006. Prediction of transporter family from protein sequence by support vector machine approach. *Proteins.* **62:** 218–231.
25. Han, L. Y., C. Z. Cai, Z. L. Ji, and Y. Z. Chen. 2005. Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity. *Virology.* **331:** 136–143.
26. Han, L. Y., C. Z. Cai, Z. L. Ji, Z. W. Cao, J. Cui, and Y. Z. Chen. 2004. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* **32:** 6437–6444.
27. Draper, D. E. 1999. Themes in RNA-protein recognition. *J. Mol. Biol.* **293:** 255–270.
28. Fierro-Monti, I., and M. B. Mathews. 2000. Proteins binding to duplexed RNA: one motif, multiple functions. *Trends Biochem. Sci.* **25:** 241–246.
29. Peculis, B. A. 2000. RNA-binding proteins: if it looks like a sn(o)RNA. *Curr. Biol.* **10:** R916–R918.
30. Perez-Canadillas, J. M., and G. Varani. 2001. Recent advances in RNA-protein recognition. *Curr. Opin. Struct. Biol.* **11:** 53–58.
31. Bairoch, A., and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.
32. Cai, C. Z., L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen. 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **31:** 3692–3697.
33. Cai, C. Z., L. Y. Han, Z. L. Ji, and Y. Z. Chen. 2004. Enzyme family classification by support vector machines. *Proteins.* **55:** 66–76.
34. Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.
35. Bock, J. R., and D. A. Gough. 2001. Predicting protein-protein interactions from primary structure. *Bioinformatics.* **17:** 455–460.
36. Ding, C. H., and I. Dubchak. 2001. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics.* **17:** 349–358.
37. Cai, Y. D., X. J. Liu, X. B. Xu, and K. C. Chou. 2002. Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **23:** 267–274.
38. Cai, Y. D., X. J. Liu, X. B. Xu, and K. C. Chou. 2002. Prediction of protein structural classes by support vector machines. *Comput. Chem.* **26:** 293–296.
39. Burges, C. J. C. 1998. A tutorial on support vector machine for pattern recognition. *Data Min. Knowl. Disc.* **2:** 121–167.
40. Veropoulos, K., C. Campbell, and N. Cristianini. 1999. Controlling the sensitivity of support vector machines. *In* Proceedings of the International Joint Conference on Artificial Intelligence (UCAI99). T. Dean, editor. Morgan Kaufmann, Stockholm, Sweden. 55–60.
41. Brown, M. P., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA.* **97:** 262–267.
42. Laity, J. H., B. M. Lee, and P. E. Wright. 2001. Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* **11:** 39–46.
43. Barrera, F. N., J. A. Poveda, J. M. Gonzalez-Ros, and J. L. Neira. 2003. Binding of the C-terminal sterile alpha motif (SAM) domain of human p73 to lipid membranes. *J. Biol. Chem.* **278:** 46878–46885.
44. Chang, S., T. ran Ma, R. D. Miranda, M. E. Balestra, R. W. Mahley, and Y. Huang. 2005. Lipid- and receptor-binding regions of apolipoprotein E4 fragments act in concert to cause mitochondrial dysfunction and neurotoxicity. *Proc. Natl. Acad. Sci. USA.* **102:** 18694–18699.
45. Chen, M. H., I. Ben-Efraim, G. Mitrousis, N. Walker-Kopp, P. J. Sims, and G. Cingolani. 2005. Phospholipid scramblase 1 contains a nonclassical nuclear localization signal with unique binding site in importin alpha. *J. Biol. Chem.* **280:** 10599–10606.
46. Vishwanathan, S. V. N., and A. J. Smola. 2003. Fast kernels for string and tree matching. Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference. S. Becker, S. Thrun, and K. Obermayer, editors. MIT Press, Cambridge, MA. AA11.
47. Ratsch, G., S. Sonnenburg, and B. Scholkopf. 2005. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics.* **21 (Suppl. 1):** 369–377.
48. Zien, A., G. Ratsch, S. Mika, B. Scholkopf, T. Lengauer, and K. R.

Muller. 2000. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics.* **16:** 799–807.
49. Jaakkola, T., M. Diekhans, and D. Haussler. 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, editors. AAAI Press, Menlo Park, CA. 149–158.
50. Tsuda, K., M. Kawanabe, G. Ratsch, S. Sonnenburg, and K. R. Muller. 2002. A new discriminative kernel from probabilistic models. *Neural Comput.* **14:** 2397–2414.
51. Liao, L., and W. S. Noble. 2003. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.* **10:** 857–868.
52. Vert, J-P., H. Saigo, and T. Akutsu. 2003. Local alignment kernels for biological sequences. *In* Kernel Methods in Computational Biology. MIT Press, Cambridge, MA. 131–154.
53. Leslie, C., R. Kuang, and E. Eskin. 2003. Inexact matching string kernels for protein classification. *In* Kernel Methods in Computational Biology. MIT Press, Cambridge, MA. 95–112.
54. Kuang, R., E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. 2005. Profile-based string kernels for remote homology detection and motif extraction. *J. Bioinform. Comput. Biol.* **3:** 527–550.
55. Han, L. Y., C. Z. Cai, S. L. Lo, M. C. Chung, and Y. Z. Chen. 2004. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA.* **10:** 355–368.
56. Lugo, M. R., and F. J. Sharom. 2005. Interaction of LDS-751 with P-glycoprotein and mapping of the location of the R drug binding site. *Biochemistry.* **44:** 643–655.
57. Hamilton, S. E., M. Recny, and L. P. Hager. 1986. Identification of the high-affinity lipid binding site in *Escherichia coli* pyruvate oxidase. *Biochemistry.* **25:** 8178–8183.
58. Raussens, V., V. Narayanaswami, E. Goormaghtigh, R. O. Ryan, and J. M. Ruysschaert. 1996. Hydrogen/deuterium exchange kinetics of apolipophorin-III in lipid-free and phospholipid-bound states. An analysis by Fourier transform infrared spectroscopy. *J. Biol. Chem.* **271:** 23089–23095.