3692–3697 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg600

SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence

C.Z. Cai[1,2], L.Y. Han[1], Z.L. Ji[1], X. Chen[1] and Y.Z. Chen[1,*]

[1]Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore and [2]Department of Applied Physics, Chongqing University, Chongqing 400044, PR China

Received February 14, 2003; Revised March 19, 2003; Accepted April 2, 2003

ABSTRACT

Prediction of protein function is of significance in studying biological processes. One approach for function prediction is to classify a protein into functional family. Support vector machine (SVM) is a useful method for such classification, which may involve proteins with diverse sequence distribution. We have developed a web-based software, SVMProt, for SVM classification of a protein into functional family from its primary sequence. SVMProt classification system is trained from representative proteins of a number of functional families and seed proteins of Pfam curated protein families. It currently covers 54 functional families and additional families will be added in the near future. The computed accuracy for protein family classification is found to be in the range of 69.1–99.6%. SVMProt shows a certain degree of capability for the classification of distantly related proteins and homologous proteins of different function and thus may be used as a protein function prediction tool that complements sequence alignment methods. SVMProt can be accessed at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

INTRODUCTION

Knowledge about protein function is essential in the understanding of biological processes (1,2). As the gap between the amount of sequence information and functional characterization widens, increasing efforts are being directed at the development of computational tools for protein function prediction (2–5). Various methods have been developed, which include sequence similarity (6–8), evolutionary analysis (9,10), structure-based approach (11), protein/gene fusion (12,13), protein interaction (14,15) and family classification by sequence clustering (16,17).

In the absence of clear sequence or structural similarities, the criteria for comparison of distantly-related proteins become increasingly difficult to formulate (17). Moreover, not all homologous proteins have analogous functions (9). The presence of a shared domain within a group of proteins does not necessarily imply that these proteins perform the same function (18). Many proteins sharing promiscuous domains (e.g. SH2, WD40, DnaJ) are known to have very different functions (12). These problems often hinder some of the clustering-based methods (16). In addition to the development of algorithms to overcome these problems (16), different approaches that combine or complement existing methods are being explored (3,9,17,19).

It is of interest to consider protein functional family classification as a method for facilitating protein function prediction, which is expected to be particularly useful in the cases described above and may thus be used as a protein function prediction tool to complement sequence alignment methods. Functional families of various proteins have been documented (20–23). A method for the classification of proteins with diverse sequence distribution is also available. A statistical learning method, support vector machines (SVM) (24), has recently been used for classification of G-protein coupled receptors (25) and DNA-binding proteins (26). It has also been employed in a number of other protein studies including protein–protein interaction prediction (15), fold recognition (27), solvent accessibility (28) and structure prediction (29,30). The prediction accuracy ranges from 65 to 91.4% in these studies. Thus SVM classification of protein functional family may be potentially developed into a protein function prediction tool to complement methods based on sequence similarity and clustering.

Instead of direct comparison or clustering of sequences, SVM classification is based on the analysis of physicochemical properties of a protein generated from its sequence (25–30). Samples of proteins known to be in a functional class (positive samples) and those not in the class (negative samples) are used to train a SVM system to recognize specific features and classify proteins into either the functional class or outside of the class. Such an approach may be applied to functional prediction for both distantly-related and closely-related proteins. Proteins

*To whom correspondence should be addressed. Tel: +65 68746877; Fax: +65 67746756; Email: yzchen@cz3.nus.edu.sg

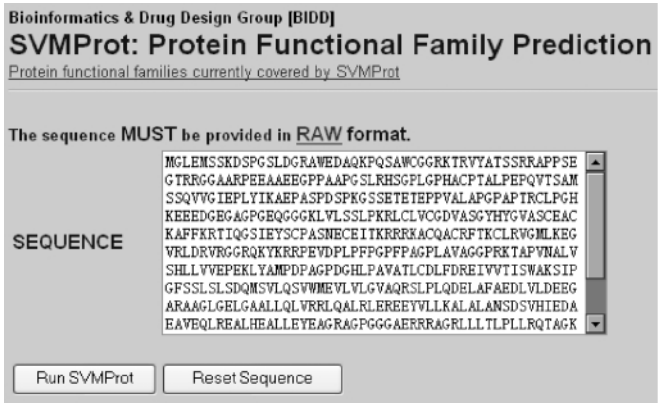

Figure 1. SVMProt web page.

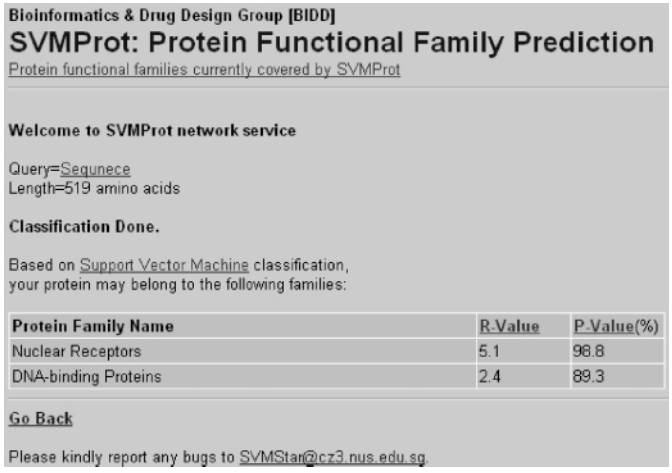

Figure 2. Example of the SVMProt output returned to the user.

of specific functional class share common structural and chemical features essential for performing similar functions (20–22). Given sufficient samples of proteins of specific function, SVM can be trained and used to recognize proteins with characteristics for a particular function (15,25,26).

We have developed a web-based software, SVMProt, for the classification of a protein into functional class from its primary sequence. The functionally distinguished classes of proteins are collected from several databases (20–23,31,32) that include all major classes of enzymes, receptors, transporters, channels, DNA-binding proteins and RNA-binding proteins. The core SVM program used in SVMProt is SVM★ which has recently been developed and tested for the classification of DNA-binding proteins (26). SVMProt is specifically trained and tested on each of the functional classes currently collected. Its usefulness on protein functional classification is evaluated. Its capability in the classification of distantly related proteins and homologous proteins of different function is also studied.

SOFTWARE ACCESS

The SVMProt web page is at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi and it is shown in Figure 1. The sequence of a protein, in RAW format and containing no non-amino acid letters, can be input in a window provided. A sequence of less than 50 amino acids is not accepted. The computed result is displayed in a separate window as shown in Figure 2. Depending on the computed result, one of the following four outcomes is displayed. If the input protein is predicted to belong to one or more functional families, then the name of each family is displayed. For some protein families, a cross-link to the respective protein family database is provided and that of more families will be added. If the input protein is predicted to not belong to any of the functional classes currently included in SVMProt, then a message of 'Your input protein is not in any of the functional classes currently covered by SVMProt' is displayed. If the input sequence contains invalid characters or abnormal composition such as a long stretch of consecutive single letters, then a message of 'invalid character . . . ' or 'your input sequence is not a valid sequence' is displayed. If the input sequence is less than 50 amino acids, then a message of 'your input sequence is less than 50 amino acids' is displayed.

METHODS

Table 1 lists the protein functional families currently covered by SVMProt. These include 46 families of enzymes from BRENDA (20), G-protein coupled receptors from GPCRDB (21), nuclear receptors from NucleaRDB (21), tyrosine receptor kinases derived from NCBI (31), five families of channels and one family of transporters from TCDB (22) and LGICdb (23) and DNA- and RNA-binding proteins derived from SWISS-PROT (32). Additional families of transporters will be added very soon. Other families of proteins are being searched and collected. The updated list of functional classes is provided in the SVMProt web page.

SVMProt is trained for protein classification in the following manner. First, every protein sequence is represented by specific feature vector assembled from encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility for each residue in the sequence (15,25–30). Three descriptors, composition (C), transition (T) and distribution (D), are used to describe global composition of each of these properties (33). C is the number of amino acids of a particular property (such as hydrophobicity) divided by the total number of amino acids. T characterizes the percent frequency with which amino acids of a particular property is 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 is located respectively.

A hypothetical protein sequence AEAAAEAEEAAAAAEA EEEAAEEAEEEAAE, as shown in Figure 3, has 16 alanines ($n1 = 16$) and 14 glutamic acids ($n2 = 14$). The composition for these two amino acids are $n1 \times 100.00/(n1 + n2) = 53.33$ 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 percent 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 $1/30 \times 100.00 = 3.33$, $5/30 \times 100.00 = 16.67$, $12/30 \times 100.00 = 40.0$, $20/30 \times 100.00 = 66.67$, $29/30 \times 100.00 = 96.67$. Likewise, the D descriptor

Table 1. List of protein families currently covered by SVMProt, statistics of datasets and prediction results. Predicted results are given in *TP* (true positive), *FN* (false negative), *TN* (true negative), *FP* (false positive), and *Q* (overall accuracy). Number of positive or negative samples in testing and independent evaluation sets is $TP + FN$ or $TN + FP$, respectively

Protein family	Training set		Testing set				Independent evaluation set				*Q* (%)
	Positive	Negative	Positive		Negative		Positive		Negative		
			TP	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*	
EC 1.1 Oxidoreductases acting on the CH-OH group of donors	383	896	743	23	1384	9	452	54	932	60	92.4
EC 1.2 Oxidoreductases acting on the aldehyde or oxo group of donors	256	1127	233	3	1156	13	200	32	972	23	95.5
EC 1.3 Oxidoreductases acting on the CH-CH group of donors	170	871	91	5	1429	2	75	33	985	15	95.7
EC 1.4 Oxidoreductases acting on the CH-NH_2 group of donors	80	459	60	3	1836	7	44	13	992	10	97.8
EC 1.5 Oxidoreductases acting on the CH-NH group of donors	129	1129	42	0	1117	3	35	33	983	21	95.0
EC 1.6 Oxidoreductases acting on NADH or NADPH	434	776	729	3	1516	15	531	42	971	33	95.2
EC 1.7 Oxidoreductases acting on other nitrogenous compounds as donors	86	1088	24	1	1224	0	36	10	1003	3	98.8
EC 1.8 Oxidoreductases acting on a sulfur group of donors	106	734	74	3	1580	2	56	30	1005	2	97.1
EC 1.9 Oxidoreductases acting on a heme group of donors	122	480	712	0	1817	0	400	18	995	5	98.4
EC 1.10 Oxidoreductases acting on diphenols and related substances as donors	48	431	23	0	1879	0	22	10	1005	0	99.0
EC 1.11 Oxidoreductases acting on a peroxide as acceptor	89	569	95	0	1740	2	73	14	997	7	98.1
EC 1.13 Oxidoreductases acting on single donors with incorporation of molecular oxygen (oxygenases)	83	721	52	1	1581	9	46	10	1001	4	98.7
EC 1.14 Oxidoreductases acting on paired donors with incorporation or reduction of molecular oxygen	201	1146	157	2	1166	3	127	24	993	13	96.8
EC 1.15 Oxidoreductases acting on superoxide as acceptor	60	1196	58	2	1119	1	54	7	1007	0	99.3
EC 1.17 Oxidoreductases acting on CH_2 groups	65	1197	58	6	1121	0	46	12	1006	2	98.7
EC 1.18 Oxidoreductases acting on iron-sulfur proteins as donors	64	814	47	1	1501	0	41	11	1006	0	99.0
EC 2.1 Transferases transferring one-carbon groups	486	1184	330	0	1103	1	287	76	920	74	88.9
EC 2.2 Transferases transferring aldehyde or ketone residues	35	1197	30	2	1121	0	26	5	1005	3	99.2
EC 2.3 Acyltransferases	302	1001	246	0	1284	4	196	44	966	27	94.2
EC 2.4 Glycosyltransferases	427	1180	264	2	1110	5	245	58	933	64	90.6
EC 2.5 Transferases transferring alkyl or aryl groups, other than methyl groups	320	1024	225	0	1284	1	197	53	964	39	92.7
EC 2.6 Transferases transferring nitrogenous groups	132	1109	79	2	1206	1	71	19	995	12	97.2
EC 2.7 Transferases transferring phosphorus-containing groups	1133	1334	1024	2	581	4	1217	195	759	202	83.3
EC 2.8 Transferases transferring sulfur-containing groups	60	541	22	1	1772	1	19	14	1003	2	98.5
EC 3.1 Hydrolases acting on ester bonds	760	1295	453	5	966	13	97	439	954	31	69.1
EC 3.2 Glycosylases	337	867	379	2	1397	13	268	49	939	51	92.3
EC 3.3 Hydrolases acting on ether bonds	54	843	29	0	1474	1	35	5	1008	0	99.5
EC 3.4 Hydrolases acting on peptide bonds (Peptidases)	436	1188	240	4	1112	3	217	59	959	43	92.0
EC 3.5 Hydrolases acting on carbon-nitrogen bonds, other than peptide bonds	414	1145	181	3	1137	2	199	73	931	60	89.5
EC 3.6 Hydrolases acting on acid anhydrides	693	1089	770	2	1196	2	646	75	951	42	93.2
EC 4.1 Carbon-carbon lyases	546	1145	776	5	1113	17	547	62	881	105	89.5
EC 4.2 Carbon-oxygen lyases	505	1231	382	1	1047	2	324	79	915	77	88.8
EC 4.3 Carbon-nitrogen lyases	96	803	86	2	1514	0	67	12	999	9	98.1
EC 4.4 Carbon-sulfur lyases	40	1194	18	11	1118	0	15	15	1004	1	98.5
EC 4.6 Phosphorus-oxygen lyases	63	989	26	0	1319	1	23	21	1002	2	97.8
EC 5.1 Racemases and Epimerases	144	830	72	0	1464	8	65	29	981	19	95.6
EC 5.2 Cis-trans-Isomerases	78	673	24	0	1643	0	32	17	1005	2	98.2
EC 5.3 Intramolecular oxidoreductases	230	950	174	2	1355	9	159	21	982	25	96.1
EC 5.4 Intramolecular transferases	144	1172	55	2	1132	7	65	26	997	7	97.0
EC 5.5 Intramolecular lyases	22	1196	14	4	1121	0	14	2	1006	1	99.7
EC 5.99 Other Isomerases	68	705	73	0	1597	7	58	8	994	9	98.4
EC 6.1 Ligases forming carbon-oxygen bonds	281	1115	381	1	1185	13	286	29	980	27	95.8
EC 6.2 Ligases forming carbon-sulfur bonds	81	947	71	0	1362	2	53	18	1001	3	98.0
EC 6.3 Ligases forming carbon-nitrogen bonds	381	1133	358	2	1148	3	294	57	946	45	92.4
EC 6.4 Ligases forming carbon-carbon bonds	48	963	26	0	1347	1	29	4	1003	1	99.5
EC 6.5 Ligases forming phosphoric ester bonds	30	1198	16	10	1095	0	18	8	979	3	98.9
G-protein coupled receptors	680	586	2694	6	1704	7	836	9	933	66	95.9
Nuclear receptors	334	538	601	7	1755	6	221	26	962	24	95.9
Tyrosine kinase receptors	14	1197	3	0	1121	0	5	2	1006	2	99.6
α-type channels	96	1037	14	0	1232	1	6	5	967	9	98.6
β-barrel porins	83	1076	19	0	1237	2	11	4	1003	5	99.1
Pore-forming toxins (proteins and peptides)	105	948	24	0	1344	0	16	12	997	0	98.8
Electrochemical potential-driven transporters porters (symporters, uniporters, antiporters)	201	450	274	4	1815	1	94	12	942	40	95.2
DNA-binding proteins	1943	1353	2308	10	799	13	1938	188	683	239	86.0
RNA-binding proteins	871	1120	610	2	1153	4	613	127	898	80	88.0

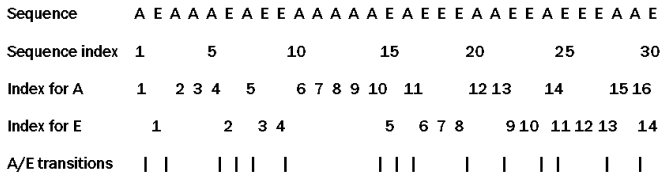

Figure 3. Hypothetical sequence for illustration of derivation of the feature vector of a protein.

for Es is 6.67, 26.67, 60.0, 76.67, 100.0. Overall, the amino acid composition descriptors for this sequence are $C = (53.33, 46.67)$, $T = (51.72)$ and $D = (3.33, 16.67, 40.0, 66.67, 96.67, 6.67, 26.67, 60.0, 76.67, 100.0)$, respectively.

Descriptors for other properties can be computed by a similar procedure and all the descriptors are combined to form the feature vector. In most studies, amino acids are divided into three classes for each property and thus the three descriptors for each property consist of 21 elements: three for C, three for T and 15 for D (15,25–30,33).

SVMProt is fed and trained with examples of proteins of a particular functional family (positive samples) and those that do not belong to this family (negative samples). The feature vectors of these positive and negative samples are input into the SVMProt system. The trained SVMProt system can then be used to classify a protein into either the positive group (protein is predicted to be in the family) or the negative group (protein is predicted to not belong to the family). Because protein feature vectors describe global composition of various physicochemical properties, SVMProt cannot address such questions as which part of a protein sequence is likely to match with a protein family.

All distinct protein members in each family found by us are used to construct positive samples for training SVMProt. More proteins are being searched which will be added in training and testing SVMProt. The negative samples for training are selected from seed proteins of the curated protein families in the Pfam database (34) excluding those that belong to the family under study. Training sets of both positive and negative samples are further screened so that only essential proteins that optimally represent each class are retained. The SVMProt training system for each family is optimized and tested by using separate testing sets of both positive and negative samples. While possible, all the remaining distinct proteins in each functional family (not in the training set of that family) are used as positive samples and all the remaining representative seed proteins in Pfam curated families are used to construct negative samples in a testing set. The performance of SVMProt classification is further evaluated by using independent sets of both positive and negative samples. There is no duplicate protein in each training, testing or independent evaluation set. The number of both positive and negative samples of proteins for the training, testing and independent evaluation sets of every functional class is given in Table 1.

The theory of SVM had been described in the literature (15,24–30). Thus only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory (24). In linearly separable cases, SVM constructs a hyperplane which separates two different groups of feature vectors with a maximum margin. A feature vector is represented by $\mathbf{x}_i$, with physicochemical descriptors of a protein as its components. The hyperplane is constructed by finding another vector $\mathbf{w}$ and a parameter b that minimizes $\| \mathbf{w} \|^2$ and satisfies the following conditions:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \quad \text{Group 1 (positive)} \qquad \mathbf{1}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \quad \text{Group 2 (negative)} \qquad \mathbf{2}$$

where 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 $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}$ can be classified by:

$$\text{sign}[(\mathbf{w} \cdot \mathbf{x}) + b] \qquad \mathbf{3}$$

In non-linearly separable cases, SVM maps the input variable into a high dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. An example of a kernel function is the Gaussian kernel which has been extensively used in different studies (15,24–30):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2} \qquad \mathbf{4}$$

Linear support vector machine is applied to this feature space and then the decision function is given by:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b \right) \qquad \mathbf{5}$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad \mathbf{6}$$

under conditions:

$$a_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \mathbf{7}$$

A positive or negative value from Eq. 3 or Eq. 5 indicates that the vector $\mathbf{x}$ belongs to the positive or negative group, respectively. To further reduce the complexity of parameter selection, hard margin SVM with threshold instead of soft margin SVM with threshold is used in SVMProt.

Scoring of SVM classification of proteins has been estimated by a reliability index and its usefulness has been demonstrated by statistical analysis (29). A slightly modified reliability score, R-value, is used in SVMProt:

$$R\text{-value} = \begin{cases} 1 & \text{if } d < 0.2 \\ \dfrac{d}{0.2} + 1 & \text{if } 0.2 \leq d < 1.8 \\ 10 & \text{if } d \geq 1.8 \end{cases} \qquad \mathbf{8}$$

where d is the distance between the position of the vector of a classified protein and the optimal separating hyperplane in the hyperspace. There is a statistical correlation between R-value

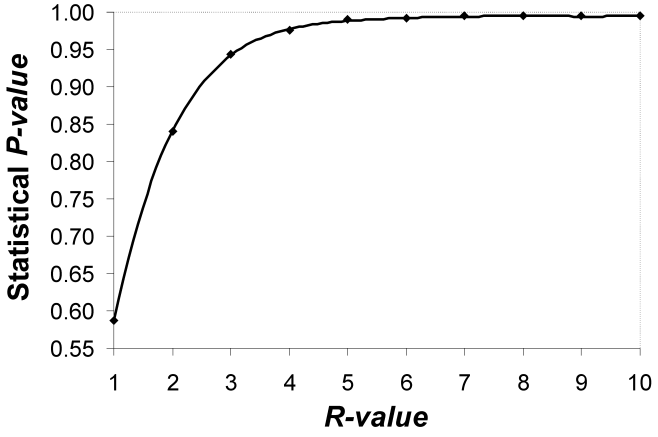

Figure 4. Statistical relationship between the *R*-value and *P*-value (probability of correct classification) derived from analysis of 9932 positive and 45 999 negative samples of proteins.

and expected classification accuracy (probability of correct classification) (29). Thus another quantity, *P*-value, is introduced to indicate the expected classification accuracy. *P*-value is derived from the statistical relationship, shown in Figure 4, between the *R*-value and actual classification accuracy based on the analysis of 9932 positive and 45 999 negative samples of proteins.

As in the case of all discriminative methods (24,35), the performance of SVMProt classification can be measured by the quantity of true positives (*TP*), true negatives (*TN*), false positives (*FP*), false negatives (*FN*) and the overall accuracy (*Q*) given below:

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad \mathbf{9}$$

RESULTS AND REMARKS

The results for the classification of each of the functional classes are given in Table 1. All the computed *TP*, *TN*, *FP*, *FN* and *Q* are given in the table. The overall accuracy *Q* of protein classification ranges from 69.1 to 99.6%, which is on average slightly improved from that obtained in other SVM studies of proteins (15,24–30). One possible reason for this improvement is the use of representative proteins of Pfam curated families as negative samples for SVM classification, which provides a more comprehensive sampling of proteins not in a functional class.

Some low sequence similarity proteins share similar function (36–38). Efforts have been directed at exploration of various novel approaches in predicting the function of these distantly related proteins (16,37,39). SVMProt is tested on 24 randomly selected distantly related proteins in seven families. Sequence similarity E-value for each of these proteins from BLAST search against most members of its family is significantly higher than the commonly accepted value of 0.05 for similarity proteins. Thus alignment methods may not work well for these proteins. Fourteen proteins are correctly classified by SVMProt, which accounts for 58.3% of all distantly related proteins studied. This suggests that, to a certain extent, SVMProt is useful for the classification of distantly related proteins.

Homologous proteins do not necessarily have analogous function (9) and there are certain levels of difficulty to distinguish them using sequence alignment methods. SVMProt is tested to four pairs of homologous proteins of different families and the results are shown in Table 2. While all eight proteins are correctly classified into their respective family, only five of them are not classified into the family of their respective homolog, representing 62.5% of all the homologous proteins examined. This limited study seems to indicate that SVMprot has a certain degree of capability for classification of homologous proteins of different functions. Further analysis is needed to provide a more objective assessment.

The ability of SVMProt in the classification of some distantly related proteins and homologous proteins of different functions probably results from the use of a combination of physicochemical properties to represent a protein. Protein function is determined by specific structural and chemical features at substrate binding sites (20). 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 structure and solvent accessibility which are used in the construction of the SVMProt feature vectors for proteins.

As shown in Table 1, there are several families with substantially high *Q* score (~90%) but relatively modest *TP*:*FN* ratio (<100:37). Generally, SVMProt gives an accurate prediction of *TN*s. The imbalance between the number of proteins in a family and those outside of the family may thus lead to cases of high *Q* score with modest *TP*:*FN* ratio. Examination of *FN* proteins of these families shows that many of these proteins either belong to more than one family or contain a domain shared by proteins in another family. These proteins are often classified into the related family. An analysis of a broad range of families indicates that a substantial portion (61.3%) of incorrectly classified proteins are of low sequence similarity to most of the other members in its family (i.e. the sequence similarity score E value of each of these proteins against most members of its family is significantly higher than 0.05). The percentage of low sequence similarity proteins in a family is not expected to be very high. Therefore, our study seems to suggest that sequence distance has a certain level of influence on the accuracy of SVM classification.

Several factors may affect the prediction accuracy. One is the diversity of protein samples. It is likely that not all possible types of proteins are adequately represented in some functional classes. This can be improved along with the availability of more protein data. SVM prediction may be further improved by using more comprehensive and refined set of protein descriptors. The SVM optimization procedure and feature vector selection algorithm may also be improved by adding additional constraints and by incorporating independent component analysis and kernel PCA in the preprocessing steps.

Our study suggests that SVM has potential in the classification of proteins into functional families. SVMProt appears to have a certain level of capability for classification of distantly related proteins and homologous proteins of different functions and, thus, potentially may be used as a protein function prediction tool that complements sequence alignment methods.

Table 2. Assessment of SVMProt classification of homologous proteins of different functions

Protein 1 (P1)	Family1 (F1)	Protein 2 (P2)	Family2 (F2)	Similarity score E-value	Classification
Glycolate oxidase (P05414)	EC1.1	IPP isomerase (Q8PW37)	EC5.3	3.00E-07	P1→F1; P2→F2
Creatinase (P38488)	EC3.5	Xaa-Pro dipeptidase (O58885)	EC3.4	3.00E-15	P1→F1; P2→F1, F2
Cystathionine gamma-synthase (P38675)	EC4.2	Methionine gamma-lyase (P13254)	EC4.4	2.00E-15	P1→F1; P2→F1, F2
Cystathionine gamma-synthase (P38676)	EC4.2	Cystathionine gamma-lyase (Q8VCN5)	EC4.4	1.00E-12	P1→F1; P2→F1, F2

P1→F1 indicates classification of protein P1 into family F1.
P2→F1, F2 indicates classification of protein P2 into both family F1 and family F2.

Further improvements on protein functional family coverage, sample collection and SVM algorithm may enable the development of SVMProt into a useful protein function prediction tool.

REFERENCES

1. Eisenberg,D., Marcotte,C.A., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
2. Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genomes and back. *J. Mol. Biol.*, **283**, 707–725.
3. Pellegrini,M. (2001) Computational methods for protein function analysis. *Curr. Opin. Chem. Biol.*, **5**, 46–50.
4. Teichman,S.A. and Mitchison,G. (2000) Computing protein function. *Nat. Biotechnol.*, **18**, 27.
5. Huynen,M., Snel,B., Lathe,W. and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
6. Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nature Genet.*, **18**, 313–318.
7. Baxevanis,A.D. (1998) Practical aspects of multiple sequence alignment. *Methods Biochem. Anal.*, **39**, 172–188.
8. Schuler,G.D. (1998) Sequence alignment and database searching. *Methods Biochem. Anal.*, **39**, 145–171.
9. Benner,S.A., Chamberlin,S.G., Liberles,D.A., Govindarajan,S. and Knecht,L. (2000) Functional inferences from reconstructed evolutionary biology involving rectified databases—an evolutionarily grounded approach to functional genomics. *Res. Microbiol.*, **151**, 97–106.
10. Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.
11. Teichmann,S.A., Murzin,A.G. and Chothia,C. (2001) Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.*, **11**, 354–363.
12. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
13. Enright,A.J., Iliopoulos,I., Kyrpides,N. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
14. Aravind,L. (2000) Guilt by association: contextual information in genome analysis. *Genome Res.*, **10**, 1074–1077.
15. Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**, 455–462.
16. Enright,A.J., Van Dongen,S.V. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
17. Enright,A.J. and Ozounis,C.A. (2000) GeneRage: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.
18. Henikoff,S., Greene,E.A., Pietrokovski,S., Bork,P., Attwood,T.K. and Hood,L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.
19. Ponting,C.P. (2001) Issues in predicting protein function from sequence. *Brief Bioinform.*, **2**, 19–29.
20. Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, **30**, 47–49.
21. Horn,F., Vriend,G. and Cohen,F.E. (2001) Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.*, **29**, 346–349.
22. Saier,M.H. Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 354–411.
23. Le Novere,N. and Changeux,J.-P. (2001) LGICdb: the ligand-gated ion channel database. *Nucleic Acids Res.*, **29**, 294–295.
24. Burges,C.J.C. (1998) A tutorial on Support Vector Machine for pattern recognition. *Data Min. Knowl. Disc.*, **2**, 121–167.
25. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
26. Cai,C.Z., Wang,W.L. and Chen,Y.Z. (2003) Support Vector Machine classification of physical and biological datasets. *Inter. J. Mod. Phys. C.*, in press.
27. Ding,C.H.Q. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.
28. Yuan,Z., Burrage,K. and Mattick,J.S. (2002) Prediction of protein solvent accessibility using support vector machines. *Proteins*, **48**, 566–570.
29. Hua,S.J. and Sun,Z.R. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.
30. Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2002) Prediction of protein structural classes by support vector machines. *Comput. Chem.*, **26**, 293–296.
31. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
32. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
33. Dubchak,I., Muchnik,I., Holbrook,S.R. and Kim,S.-H. (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl Acad. Sci. USA*, **92**, 8700–8704.
34. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
35. Baldi,P., Brunak,S., Chauvin,Y., Anderson,C.A.F. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–419.
36. Nagano,N., Porter,C.T. and Thornton,J.M. (2001) The (betaalpha)(8) glycosidases: sequence and structure analyses suggest distant evolutionary relationships. *Protein Eng.*, **14**, 845–855.
37. Frishman,D. and Argos,P. (1992) Recognition of distantly related protein sequences using conserved motifs and neural networks. *J. Mol. Biol.*, **228**, 951–962.
38. Miyata,Y. and Nishida,E. (1999) Distantly related cousins of MAP kinase: biochemical properties and possible physiological functions. *Biochem. Biophys. Res. Commun.*, **266**, 291–295.
39. Yang,A.S. (2002) Structure-dependent sequence alignment for remotely related proteins. *Bioinformatics*, **18**, 1658–1665.