Enzyme Family Classification by Support Vector Machines

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

[1]*Department of Applied Physics, Chongqing University, Chongqing, Peoples Republic of China*

[2]*Department of Computational Science, National University of Singapore, Singapore*

ABSTRACT **One approach for facilitating protein function prediction is to classify proteins into functional families. Recent studies on the classification of G-protein coupled receptors and other proteins suggest that a statistical learning method, Support vector machines (SVM), may be potentially useful for protein classification into functional families. In this work, SVM is applied and tested on the classification of enzymes into functional families defined by the Enzyme Nomenclature Committee of IUBMB. SVM classification system for each family is trained from representative enzymes of that family and seed proteins of Pfam curated protein families. The classification accuracy for enzymes from 46 families and for non-enzymes is in the range of 50.0% to 95.7% and 79.0% to 100% respectively. The corresponding Matthews correlation coefficient is in the range of 54.1% to 96.1%. Moreover, 80.3% of the 8,291 correctly classified enzymes are uniquely classified into a specific enzyme family by using a scoring function, indicating that SVM may have certain level of unique prediction capability. Testing results also suggest that SVM in some cases is capable of classification of distantly related enzymes and homologous enzymes of different functions. Effort is being made to use a more comprehensive set of enzymes as training sets and to incorporate multiclass SVM classification systems to further enhance the unique prediction accuracy. Our results suggest the potential of SVM for enzyme family classification and for facilitating protein function prediction. Our software is accessible at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi. Proteins 2004;55:66–76.**

© 2004 Wiley-Liss, Inc.

Key words: classification; enzyme; support vector machine; protein family; protein function; protein function prediction; protein sequence

INTRODUCTION

Determination of protein function is essential for understanding biological processes.[1,2] Computational tools for protein function prediction have been developed[1,3–5] using a variety of methods including sequence similarity,[6–8] evolutionary analysis,[9,10] hidden Markov models,[11] structural consideration,[12,13] protein/gene fusion,[14,15] protein interaction,[16,17] motifs,[18] neural-networks,[11,19] and family classification by sequence clustering.[20,21] In the absence of clear sequence or structural similarities, the criteria for comparison of distantly-related proteins become increasingly difficult to formulate.[20] Moreover, not all homologous proteins have analogous functions.[10] The presence of shared domain within a group of proteins does not necessarily imply that these proteins perform the same function.[22] Many proteins sharing promiscuous domains are known to have very different functions.[15] These problems have prompted effort and interest in developing new clustering algorithms[21] and exploring novel approaches that combine or complement existing methods.[5,10,20,23]

One approach for facilitating protein function prediction is to classify proteins into functional families. 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] from their primary sequences, both families contain proteins of diverse sequence distributions. Moreover, SVM has been used in a number of other protein studies including prediction of protein-protein interaction,[17] fold recognition,[27,28] study of solvent accessibility[29] and structure prediction.[30,31] The prediction accuracy derived from these studies ranges from 65% to 91.4%, suggesting the potential of SVM in facilitating the study of various protein classification problems. Because of its ability in classifying proteins of diverse sequences, SVM is expected to be particularly useful for the classification of distantly related proteins and it can thus be used to complement sequence similarity and clustering methods.

Instead of direct comparison or clustering of sequences, SVM classification is based on the analysis of physicochemical properties of a protein derived from its primary sequence.[25–27,29–31] Samples of proteins known to be in a class (positive samples) and those not in the class (negative samples) are used to train a SVM classification system to recognize specific features and classify proteins either into the class or outside the class. Such an approach may be applied to classification of both distantly-related proteins and other proteins into their respective functional families. Proteins of specific functional family share common structural and chemical features essential for performing similar functions.[32] Given sufficient samples of proteins of specific function, SVM may be trained and used to

*Correspondence to: Y.Z. Chen, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore. E-mail: yzchen@cz3.nus.edu.sg

Received 29 May 2003; Accepted 24 October 2003

© 2004 WILEY-LISS, INC.

TABLE I. List of Enzyme Families Studied in this Work, Statistics of Datasets and Prediction Results[†]

	Training set		Testing set				Independent evaluation set						
			Positive		Negative		Positive		Negative		Q_p	Q_n	
Enzyme family (EC number)	Positive	Negative	*TP*	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*	(%)	(%)	*C*
Oxidoreductases acting on the CH—OH group of donors (EC 1.1)	383	896	743	23	1384	9	452	54	932	60	89.3	94.0	0.830
Oxidoreductases acting on the aldehyde or oxo group of donors (EC 1.2)	256	1127	233	3	1156	13	200	32	972	23	86.2	97.7	0.852
Oxidoreductases acting on the CH—CH group of donors (EC 1.3)	170	871	91	5	1429	2	75	33	985	15	69.4	98.5	0.738
Oxidoreductases acting on the CH—NH_2 group of donors (EC 1.4)	80	459	60	3	1836	7	44	13	992	10	77.2	99.0	0.782
Oxidoreductases acting on the CH—NH group of donors (EC 1.5)	129	1129	42	0	1117	3	35	33	983	21	51.5	97.9	0.541
Oxidoreductases acting on NADH or NADPH (EC 1.6)	434	776	729	3	1516	15	531	42	971	33	92.7	96.7	0.897
Oxidoreductases acting on other nitrogenous compounds as donors (EC 1.7)	86	1088	24	1	1224	0	36	10	1003	3	78.3	99.7	0.844
Oxidoreductases acting on a sulfur group of donors (EC 1.8)	106	734	74	3	1580	2	56	30	1005	2	65.1	99.8	0.780
Oxidoreductases acting on a heme group of donors (EC 1.9)	122	480	712	0	1817	0	400	18	995	5	95.7	99.5	0.961
Oxidoreductases acting on diphenols and related substances as donors (EC 1.10)	48	431	23	0	1879	0	22	10	1005	0	68.8	100	0.825
Oxidoreductases acting on a peroxide as acceptor (EC 1.11)	89	569	95	0	1740	2	73	14	997	7	83.9	99.3	0.865
Oxidoreductases acting on single donors with incorporation of molecular oxygen (oxygenases) (EC 1.13)	83	721	52	1	1581	9	46	10	1001	4	82.1	99.6	0.863
Oxidoreductases acting on paired donors, with incorporation or reduction of molecular oxygen (EC 1.14)	201	1146	157	2	1166	3	127	24	993	13	84.1	98.7	0.855
Oxidoreductases acting on superoxide as acceptor (EC 1.15)	60	1196	58	2	1119	1	54	7	1007	0	88.5	100	0.938
Oxidoreductases acting on CH_2 groups (EC 1.17)	65	1197	58	6	1121	0	46	12	1006	2	79.3	99.8	0.865
Oxidoreductases acting on iron-sulfur proteins as donors (EC 1.18)	64	814	47	1	1501	0	41	11	1006	0	78.8	100	0.883
Transferases transferring one-carbon groups (EC 2.1)	486	1184	330	0	1103	1	287	76	920	74	79.1	92.6	0.717
Transferases transferring aldehyde or ketone residues (EC 2.2)													
Acyltransferases (EC 2.3)	302	1001	246	0	1284	4	196	44	966	27	81.7	97.3	0.812
Glycosyltransferases (EC 2.4)	427	1180	264	2	1110	5	245	58	933	64	80.9	93.6	0.739

TABLE I. (Continued)

	Training set		Testing set				Independent evaluation set						
			Positive		Negative		Positive		Negative		Q_p	Q_n	
Enzyme family (EC number)	Positive	Negative	*TP*	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*	(%)	(%)	*C*
Transferases transferring alkyl or aryl groups, other than methyl groups (EC 2.5)	320	1024	225	0	1284	1	197	53	964	39	78.8	96.1	0.766
Transferases transferring nitrogenous groups (EC 2.6)	132	1109	79	2	1206	1	71	19	995	12	78.9	98.8	0.806
Transferases transferring phosphorus-containing groups (EC 2.7)	1133	1334	1024	2	581	4	1217	195	759	202	86.2	79.0	0.652
Transferases transferring sulfur-containing groups (EC 2.8)	60	541	22	1	1772	1	19	14	1003	2	57.6	99.8	0.715
Hydrolases acting on ester bonds (EC 3.1)	760	1295	453	5	966	13	381	155	892	93	71.1	90.6	0.636
Glycosylases (EC 3.2)	337	867	379	2	1397	13	268	49	939	51	84.5	94.8	0.792
Hydrolases acting on ether bonds (EC 3.3)	54	843	29	0	1474	1	35	5	1008	0	87.5	100	0.933
Hydrolases acting on peptide bonds (peptidases) (EC 3.4)	436	1188	240	4	1112	3	217	59	959	43	78.6	95.7	0.760
Hydrolases acting on carbon-nitrogen bonds, other than peptide bonds (EC 3.5)	414	1145	181	3	1137	2	199	73	931	60	73.2	93.9	0.683
Hydrolases acting on acid anhydrides (EC 3.6)	693	1089	770	2	1196	2	646	75	951	42	89.6	95.8	0.860
Carbon-carbon lyases (EC 4.1)	546	1145	776	5	1113	17	547	62	881	105	89.8	89.4	0.782
Carbon-oxygen lyases (EC 4.2)	505	1231	382	1	1047	2	324	79	915	77	80.4	92.2	0.727
Carbon-nitrogen lyases (EC 4.3)	96	803	86	2	1514	0	67	12	999	9	84.8	99.1	0.854
Carbon-sulfur lyases (EC 4.4)	40	1194	18	11	1118	0	15	15	1004	1	50.0	99.9	0.679
Phosphorus-oxygen lyases (EC 4.6)	63	989	26	0	1319	1	23	21	1002	2	52.3	99.8	0.684
Racemases and epimerases (EC 5.1)	144	830	72	0	1464	8	65	29	981	19	69.1	98.1	0.708
Cis-trans-isomerases (EC 5.2)	78	673	24	0	1643	0	32	17	1005	2	65.3	99.8	0.776
Intramolecular oxidoreductases (EC 5.3)	230	950	174	2	1355	9	159	21	982	25	88.3	97.5	0.851
Intramolecular transferases (EC 5.4)	144	1172	55	2	1132	7	65	26	997	7	71.4	99.3	0.788
Intramolecular lysases (EC 5.5)	22	1196	14	4	1121	0	14	2	1006	1	87.5	99.9	0.902
Other isomerases (EC 5.99)	68	705	73	0	1597	7	58	8	994	9	87.9	99.1	0.864
Ligases forming carbon-oxygen bonds (EC 6.1)	281	1115	381	1	1185	13	286	29	980	27	90.8	97.3	0.883
Ligases forming carbon-sulfur bonds (EC 6.2)	81	947	71	0	1362	2	53	18	1001	3	74.6	99.7	0.831
Ligases forming carbon-nitrogen bonds (EC 6.3)	381	1133	358	2	1148	3	294	57	946	45	83.8	95.5	0.801
Ligases forming carbon-carbon bonds (EC 6.4)	48	963	26	0	1347	1	29	4	1003	1	87.9	99.9	0.919
Ligases forming phosphoric ester bonds (EC 6.5)	30	1198	16	10	1095	0	18	8	979	3	69.2	99.7	0.765

[†] The results are given in *TP* (true positive), *FN* (false negative), *TN* (true negative), *FP* (false positive), Q_p and Q_n (Unique accuracy for prediction of positive and negative samples), *C* (Matthews correlation coefficient). Number of positive or negative samples in testing and independent evaluation sets is *TP* + *FN* or *TN* + *FP* respectively. Updated results are given at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

recognize proteins possessing characteristics of a particular function.[17, 25, 26]

In this work, the usefulness of SVM for classification of proteins into functional families is tested on enzymes from 46 enzyme families. Enzymes represent the largest and most diverse group of all proteins, catalyzing chemical reactions in the metabolism of all organisms. Enzymes are well classified into functional families according to the recommendation by the classification of enzyme nomenclature committee of IUBMB.[33] Therefore enzymes are ideal for comprehensive testing of the capability of SVM classification systems. SVM is also evaluated for its capability in the classification of distantly related enzymes and homologous enzymes of different function.

METHODS

Enzyme families are obtained from BRENDA database.[32] There are 46 enzyme families found to have substantial number of enzymes in Swiss-Prot database.[34] Sufficient number of samples is needed to train a SVM classification system for accurate classification, thus only these 46 families are studied in this work. Table I gives the list of enzyme families together with the number of enzymes in each family used for training, testing, and evaluating SVM classification system for that family.

All distinct members in each enzyme family found in Swiss-Prot database[34] are used to construct positive samples for training SVM. The negative samples corresponding to each enzyme family are selected from seed proteins of the curated protein families in the Pfam database.[35] Those seed proteins known to not belong to the enzyme family under study are used as negative samples for that family. Negative samples of each family include representative enzymes in all the other enzyme families and non-enzyme proteins such as receptors, transporters, channels, and other non-enzyme proteins. An example of the composition of negative samples in an enzyme family EC2.7 is given in Table II. There are cases such that particular proteins can be positive for more than one family and these are only included in the respective positive training set and excluded in the negative training set. Also, the EC number of some enzymes may not be specified at the time of our data collection, some of which may be tentatively included in the negative training set.

In most cases, there are multiple entries in the Swiss-Prot database for each distinct protein in each enzyme family. Thus, after the selection of the training set for a family, there is a sufficient number of entries left in Swiss-Prot database for construction of separate sets of both positive and negative samples for that family. This allows one to optimize and test the SVM training system for each family by using separate testing sets and to evaluate the prediction results by using independent evaluation sets of both positive and negative samples. While possible, all the remaining distinct enzymes in each family (not in its training set) 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 and an independent evaluation set. For proteins that belong to more than one families, they are only included in the positive training, testing, and independent evaluation set of a particular family under study. No duplicate enzyme is used in the training, testing, and independent evaluation set for each family.

Training sets of both positive and negative samples can be optimized by exchanging the incorrectly classified samples in the corresponding testing sets with non-support-vector samples in the training sets so that all the essential proteins that optimally represent each family are retained in the training sets. These essential proteins carry distinct structural and physicochemical features important to characterize the members of each family and those outside the family. The support vectors of the positive and negative samples for that family are generated from these proteins.

Prediction accuracies of statistical learning methods are typically evaluated by methods such as n-fold cross validation.[27] Our SVM system is trained by using optimized training sets which include all the essential proteins in a family. In an n-fold cross validation study, it is difficult to keep all these essential proteins within a training set. Thus in this work, evaluation of prediction accuracy is conducted by using independent evaluation sets. As will be presented in the results and discussion section of this paper, the derived prediction accuracies from our method

TABLE II. Composition of the Negative Samples for EC2.7 Family[†]

Family	Number of entries
EC 1.1	10
EC 1.2	3
EC 1.3	17
EC 1.4	6
EC 1.5	2
EC 1.6	7
EC 1.7	2
EC 1.8	1
EC 1.9	24
EC 1.10	8
EC 1.11	4
EC 1.13	4
EC 1.14	1
EC 1.15	3
EC 1.18	2
EC 2.1	11
EC 2.3	20
EC 2.4	20
EC 2.5	4
EC 3.1	30
EC 3.2	33
EC 3.3	2
EC 3.4	12
EC 3.5	9
EC 3.6	33
EC 4.1	28
EC 4.2	18
EC 4.4	7
EC 4.6	5
EC 5.1	7
EC 5.4	3
EC 5.5	1
EC 5.99	9
EC 6.1	1
EC 6.2	1
EC 6.3	20
EC 6.4	6
EC 6.5	9
Receptors	17
Transporters	53
Channels	11
Other proteins	1455

[†] Here "other proteins" include proteins know to not belong to any of the families listed and those enzymes whose EC number is not specified at the time of our data-collection.

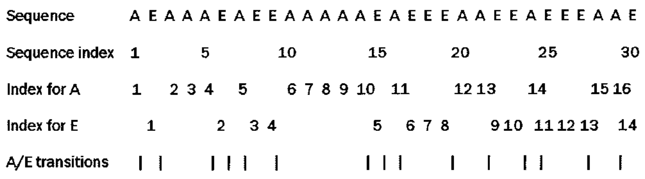
Sequence A E A A A E A E E A A A A A E A E E E A A E E A E E E A A E
Sequence index 1 5 10 15 20 25 30
Index for A 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Index for E 1 2 3 4 5 6 7 8 9 10 11 12 13 14
A/E transitions | | | | | | | | | | | | | | |

Fig. 1. Hypothetical sequence for illustration of derivation of the feature vector of a protein.

is similar to those derived from 10-fold cross validation study.

Every enzyme sequence is represented by specific feature vectors 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.[17,25–27,29–31] There is some level of overlap in the descriptors for hydrophobicity, polarity, and surface tension, which may be reduced by principle component analysis (PCA). Our own study suggests that the use of the PCA-reduced descriptors only moderately improves the accuracy for some of the families. It is thus unclear to which extent this overlap affects the accuracy of SVM classification. It is noted that reasonably accurate results have been obtained in various protein classification studies using these overlapping descriptors.[17,25–27,29–31]

Three descriptors, composition (C), transition (T), and distribution (D), are used to describe global composition of each of the properties described above.[36,37] 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 AEAAAEAEEAAAAAEAEEEAAEEAEEEAAE, as shown in Figure 1, 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 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 of a protein. 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: 3 for C, 3 for T, and 15 for D.[17,25–27,29,30,36,37]

The constructed feature vectors of both positive samples (examples of enzymes in a particular family) and negative samples (those do not belong to a particular family) are then input into SVM classification system to train it to identify features that separate positive and negative samples. The trained SVM systems can thus be used to classify an enzyme into either the positive group or the negative group of each family. This enzyme is predicted to be a member of a family if it is classified into the positive group of that family. Likewise, it is predicted to not belong to a family if it is classified into the negative group of that family. The theory of SVM has been described in the literature.[17,24–27,29–31] 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:

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

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

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 $\mathbf{x}_i$ can be classified by:

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

In nonlinearly 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:[17,24–27,29–31]

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

Based on earlier study[27,38] and our own analysis, Gaussian kernel function seems to produce better results than other kernel functions. Linear support vector machine is applied to this feature space and then the decision function is given by:

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

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

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

under conditions:

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

A positive or negative value from Eq. (3) or Eq. (5) indicates that the vector **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[39] with threshold is used. We have developed our own SVM program SVM★[26] using the sequential minimal optimization (SMO) algorithm,[40] RBF kernel and parameters C→∞ (for hard margin) and σ value of 5–35 for different enzyme families. RBF kernel is used because it has been commonly used in other SVM protein studies with consistently better performance than other kernels such as linear and polynomial.[27,38] Our own analysis on enzyme family classification suggests that the prediction accuracy using RBF kernel is at least 5% more than that using polynomial kernel.

As in the case of all discriminative methods,[24,41] the performance of SVM classification can be measured by the quantity of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*). Because the number of positive and negative samples for each family is imbalanced, two accuracies Q_p and Q_n are introduced to measure the accuracy of positive prediction (proteins belong to an enzyme family) and negative prediction (proteins do not belong to an enzyme family):

$$Q_p = \frac{TP}{TP + FN}$$

$$Q_n = \frac{TN}{TN + FP} \tag{8}$$

Another quantity suitable for evaluating the classification accuracy of imbalanced positive and negative samples is the Matthews correlation coefficient C,[42] which is given by:

$$C = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{9}$$

RESULTS AND DISCUSSION

Assessment of Overall Accuracy of SVM Enzyme Family Classification

The results for the classification of the 46 enzyme families are given in Table I. All the computed *TP*, *TN*, *FP*, and *FN* for the testing sets and independent evaluation sets of these families are given in the Table. Table I also gives the classification accuracies Q_p and Q_n and Matthews correlation coefficient C for every family measured by using independent evaluation sets. The computed Q_p, Q_n and C for the 46 enzyme families are in the range of 50.0% to 95.7%, 79.0% to 100%, and 54.1% to 96.1% respectively. These numbers on average are slightly improved from that obtained in other SVM studies of proteins.[17,24–27,29–31] 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 belonging to an enzyme family.

Table III gives a list of a number of randomly selected enzyme entries from Swiss-Prot database[34] that are not correctly classified into the corresponding family by SVM★. Amino acid sequence of each of these enzyme entries is examined to determine whether or not the classification error is caused by sequence-related problems such as fragment, incomplete chain, and mutations. As shown in Table II, these sequence-related problems do not appear to be a significant factor for the classification error.

BLAST sequence alignment of each of these enzymes against other members of its family suggests that a substantial portion (61.3%) of these incorrectly classified enzymes 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 enzymes 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 certain level of influence on the accuracy of SVM classification.

Several other factors may also affect the classification accuracy. One is the sequence diversity of protein samples in a functional family. 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 a more comprehensive and refined set of protein descriptors. 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.

The quality of our SVM classification system of a particular enzyme family can be further assessed by means of direct two-way tests. For such a purpose, a set of 3000 enzymes in a randomly selected enzyme family EC1.6 is used for testing the accuracy of positive classification for that family. It is found that 76.8% of these enzymes are correctly classified into the EC1.6 family by our SVM system. A set of 2850 randomly selected non-enzyme proteins is used for assessing the accuracy of negative classification for that enzyme family. It is found that 98.5% of these non-enzyme proteins are correctly classified as not belonging to the EC1.6 family.

Comparison Between Results From our Evaluation Method and Those of 10-Fold Cross Validation

In this work, independent evaluation sets are used to determine the accuracy of enzyme family classification. To examine whether it can provide sufficiently accurate assessment of prediction accuracy, the results from three randomly selected families using our evaluation

TABLE III. Randomly Selected Enzyme Entries From Swiss-Prot Database Which are Not Correctly Classified Into the Corresponding Family by SVM*†

EC Family number	Swiss Prot AC number	Protein name	Sequence feature	Sequence similarity to other members of family
EC 1.1	Q8YH79	Alcohol dehydrogenase	C	L
EC 1.14	P79078	Delta-9 fatty acid desaturase	C	S
EC 1.14	Q8TE42	Truncated steroid 21-hydroxylase	IC	L
EC 1.14	P14791	Heme oxygenase	C	L
EC 1.2	O67724	N-acetyl-γ-glutamyl-phosphate reductase	C	L
EC 1.2	Q57658	Aspartate-semialdehyde dehydrogenase	C	L
EC 2.1	Q9ZE37	tRNA (Guanine-N(1)-)-methyltransferase	C	S
EC 2.1	Q9PJ28	Methionyl-tRNA formyltransferase	C	S
EC 2.1	Q9UX08	Aspartate carbamoyltransferase	C	L
EC 2.1	P96111	PyrBI protein	C	L
EC 2.7	Q9JR61	Phosphatidylserine synthase	C	L
EC 2.7	Q9ZE96	Phosphatidylglycerophosphate synthase	C	L
EC 3.1	Q62087	Serum paraoxonase/arylesterase 3	C	L
EC 3.1	Q97VT7	Aryldialkylphosphatase, putative	C	S
EC 3.2	Q9EVP3	Stx2fA protein subunit	C, subunit	L
EC 3.2	Q9S9E4	rRNA-glycosidase	C	L
EC 3.2	Q41216	Trichosanthin	C	L
EC 3.5	P32320	Cytidine deaminase	C, subunit	L
EC 3.5	Q01432	AMP deaminase 3	C, subunit	L
EC 3.5	Q49135	Methenyltetrahydrofolate cyclohydrolase	C, subunit	S
EC 4.2	P73715	Endonuclease III	C	S
EC 4.2	Q8RI68	Cystathionine gamma-synthase	C	S
EC 4.3	Q8XMJ8	Argininosuccinate lyase	C	S
EC 5.1	Q980W1	UDP-glucose 4-epimerase	C	S
EC 5.1	P21955	Aldose 1-epimerase	C	L
EC 5.3	P29954	Mannose-6-phosphate isomerase	C	S
EC 5.4	Q8Z8D7	UDP-galactopyranose mutase	C	S
EC 6.1	Q8YH72	Alanyl-tRNA synthetase	C	L
EC 6.1	Q9ZDF8	Lysyl-tRNA synthetase	C	L
EC 6.1	Q9HJM5	Glutamyl-tRNA synthetase	C	L
EC 6.1	Q55486	Arginyl-tRNA synthetase	C	L
EC 6.3	P57245	Carbamoyl-phosphate synthase, small chain	C, chain	S

[†] C—Complete sequence; IC—Incomplete sequence; C, subunit—Complete sequence of subunit; C, chain—Complete sequence of chain; L—Low sequence similarity to other enzymes in a particular family; S—Significant sequence similarity to other enzymes in a particular family.

TABLE IV. Ten-fold Cross Validation Results of EC1.9 Family†

			Testing set						
	Training set		Positive		Negative		Evaluation		
Fold number	Positive	Negative	*TP*	*FN*	*TN*	*FP*	Q_p (%)	Q_n (%)	*C*
1	1127	2967	119	6	327	3	95.2	99.1	0.950
2	1127	2967	119	6	328	2	95.2	99.4	0.955
3	1126	2968	119	7	325	4	94.4	98.7	0.939
4	1127	2967	116	9	330	0	92.8	100	0.950
5	1127	2967	122	3	327	3	97.6	99.1	0.967
6	1127	2967	115	10	330	0	92.0	100	0.945
7	1126	2968	121	5	328	1	96.0	99.7	0.967
8	1126	2968	117	9	326	3	92.8	99.1	0.933
9	1127	2967	113	12	327	3	90.4	99.1	0.916
10	1127	2967	120	5	326	4	96.0	98.7	0.950
Average							94.2	99.3	0.947
Our method							95.7	99.5	0.961

[†] The result from our method is included for comparison.

method are compared with those from a 10-fold cross validation study. Table IV, Table V, and Table VI give the results of the 10-fold cross validation study for the EC1.9, EC4.4, and EC5.2 family respectively. For comparison, the results from our study are also included in the respective Table. It is found that the computed Q_p,

TABLE V. Ten-fold Cross Validation Results of EC4.4 Family[†]

	Training set		Testing set				Evaluation		
			Positive		Negative				
Fold number	Positive	Negative	*TP*	*FN*	*TN*	*FP*	Q_p (%)	Q_n (%)	*C*
1	89	2985	5	5	332	0	50.0	100	0.701
2	90	2894	7	2	333	0	77.7	100	0.879
3	89	2985	6	4	332	0	60.0	100	0.769
4	89	2985	6	4	331	1	60.0	99.6	0.710
5	89	2985	6	4	332	0	60.0	100	0.769
6	89	2985	5	5	332	0	50.0	100	0.701
7	89	2986	8	2	331	0	80.0	100	0.891
8	89	2986	5	5	331	0	50.0	100	0.701
9	89	2986	8	2	331	0	80.0	100	0.891
10	89	2986	9	1	330	1	90.0	99.6	0.897
Average							65.7	99.9	0.791
Our method							50.0	99.9	0.679

[†] The result from our method is included for comparison.

TABLE VI. Ten-fold Cross Validation Results of EC5.2 Family[†]

	Training set		Testing set				Evaluation		
			Positive		Negative				
Fold number	Positive	Negative	*TP*	*FN*	*TN*	*FP*	Q_p (%)	Q_n (%)	*C*
1	136	2990	11	4	333	0	73.3	100	0.851
2	136	2990	12	3	333	0	80.0	100	0.890
3	137	2989	9	5	334	0	64.2	100	0.795
4	137	2989	9	5	334	0	64.2	100	0.795
5	137	2990	8	6	333	0	57.1	100	0.749
6	136	2991	7	8	332	0	46.7	100	0.675
7	134	2993	11	6	330	0	64.7	100	0.797
8	134	2993	12	5	330	0	70.5	100	0.833
9	136	2991	10	5	331	1	66.7	99.7	0.770
10	136	2991	12	3	331	1	80.0	99.7	0.853
Average							66.7	99.9	0.800
Our method							65.3	99.8	0.776

[†] The result from our method is included for comparison.

Q_n, and *C* for each of these families using our method is roughly similar to those obtained by using 10-fold cross validation study. This suggests that our method may be used to assess the quality of SVM enzyme family classification, with a similar level of accuracy as that of n-fold cross validation study.

Classification of Distantly Related Enzymes

Certain proteins with very low sequence similarity to each other are known to have similar function.[20,43–45] The low sequence similarity nature of these distantly related proteins makes it difficult to use conventional sequence alignment and clustering methods. It has thus prompted the introduction of novel approaches for functional prediction of distantly related proteins. These include neural network analysis of conserved motifs,[43] energy analysis,[46] structure-dependent sequence alignment,[47] and sequence clustering-based family classification using pre-computed sequence similarity information.[21]

In this work 24 randomly selected distantly related enzymes in seven different families, shown in Table VII, are used to test the capability of SVM classification of distantly related enzymes. These include two aminotrasferases from EC2.6 family,[48] three kinases from EC2.7 family,[44,49] eight glycosyl hydrolases from EC3.2 family,[45] three proteases from EC3.4 family,[50–52] and eight enzymes from EC2.1, 3.5 and 6.1 families. Sequence similarity score E value for each of these enzymes from BLAST search against most members of its family is significantly higher than 0.05, the commonly accepted threshold for similarity proteins. Fourteen (14) enzymes are correctly classified, which accounts for 58.3% of all distantly related enzymes studied. This suggests that, to a certain extent, SVM can be used for classification of distantly related enzymes.

The ability of SVM in classification of some distantly related enzymes likely results from the use of a combination of physicochemical properties to represent an enzyme. In some cases, enzyme function is determined by specific structural and chemical features at substrate binding sites, and these features are shared by distantly related as well as other enzymes of the same family.[32]

TABLE VII. Assessment of SVM★ Classification of Distantly Related Enzymes

Classification of distantly related enzymes	Swiss-Prot AC number	Family	Correctly classified by SVM
PyrBlprotein (EC 2.1.3.2)	P96111	EC 2.1	No
Alanine aminotransferase (EC 2.6.1.2)	P24298	EC 2.6	Yes
Histidinol-phosphate aminotransferase 2 (EC 2.6.1.9)	Q8Y0Y8	EC 2.6	Yes
Casein kinase I homolog cki1 (EC 2.7.1.−)	P40233	EC 2.7	No
MUK (EC 2.7.1.37)	Q63796	EC 2.7	Yes
PRP4 kinase (EC 2.7.1.37)	Q13523	EC 2.7	No
6-phospho-β-glucosidase (EC 3.2.1.86)	Q46130	EC 3.2	Yes
β-galactosidase I (EC 3.2.1.23)	P19668	EC 3.2	Yes
β-mannanase/endoglucanase A precursor (EC 3.2.1.78)	P22533	EC 3.2	Yes
Cellulose-growth-specific protein precursor (EC 3.2.1.4)	Q00023	EC 3.2	Yes
Chitinase 1 precursor (EC 3.2.1.14)	P46876	EC 3.2	No
Endo-1, 4-β-xylanase C precursor (EC 3.2.1.8)	P26220	EC 3.2	Yes
Endoglucanase A precursor (EC 3.2.1.4)	P29719	EC 3.2	Yes
Mannosyl-oligosaccharide glucosidase (EC 3.2.1.106)	Q13724	EC 3.2	Yes
Botulinum neurotoxin type A Precursor (EC 3.4.24.69)	P10845	EC 3.4	No
Methionine aminopeptidase (EC 3.4.11.18)	O58362	EC 3.4	Yes
Xaa-Pro aminopeptidase 2 [Precursor](EC 3.4.11.9)	O43895	EC 3.4	Yes
Allantoinase (EC 3.5.2.5)	P40757	EC 3.5	No
Dihydropyrimidinase (EC 3.5.2.2)	Q14117	EC 3.5	Yes
Urea amidohydrolase (EC 3.5.1.5)	P94669	EC 3.5	Yes
Alanyl-tRNA synthetase (EC 6.1.1.7)	Q8YH72	EC 6.1	No
Lysyl-tRNA synthetase (EC 6.1.1.6)	Q9ZDF8	EC 6.1	No
Arginyl-tRNA synthetase (EC 6.1.1.19)	Q55486	EC 6.1	No
Glutamyl-tRNA synthetase (EC 6.1.1.17)	Q9HJM5	EC 6.1	No

TABLE VIII. Assessment of SVM★ Classification of Homologous Enzymes of Different Functions

Enzyme 1(E1)	Family 1 (F1)	Enzyme 2(E2)	Family 2 (F2)	Similarity Score E-Value	Classification
Glycolate oxidase(P05414)	1.1	IPP isomerase(Q8PW37)	5.3	3.00E-07	E1→F1; E2→F2
Creatinase(P38488)	3.5	Xaa-Pro dipeptidase(O58885)	3.4	3.00E–15	E1→F1; E2→F1, F2
Cystathionine gamma-synthase(P38675)	4.2	Methionine gamma-lyase(P13254)	4.4	2.00E–15	E1→F1; E2→F1, F2
Cystathionine gamma-synthase(P38676)	4.2	Cystathionine gamma-lyase(Q8VCN5)	4.4	1.00E–12	E1→F1; E2→F1, F2

E1→F1 indicates classification of enzyme E1 into family F1.
E2→F1, F2 indicates classification of enzyme E2 into both family F1 and family F2.

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,[53,54] secondary structure and solvent accessibility which are used in the construction of the SVM★ feature vectors for the enzymes. It is thus expected that, upon proper training with sufficiently diverse set of enzymes, SVM★ may be potentially used for the classification of certain types of distantly related enzymes that share common structural and chemical features.

Not all distantly related proteins of the same function have similar structural and chemical features. There are cases in which different functional groups, un-conserved with respect to position in the primary sequence, mediate the same mechanistic role, due to the flexibility at the active site.[55] This plasticity is unlikely to be sufficiently described by the physicochemical descriptors used in SVM★. Therefore SVM★ in the present form is not expected to be capable of classification of these types of distantly related enzymes.

Classification of Homologous Enzymes of Different Functions

Homologous proteins not necessarily have analogous function.[10] It is thus useful to develop protein function prediction methods that can distinguish homologous proteins of different functions. 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.[56] While these changes appear to be small at the local sequence level, some of the aspects of these changes may be reflected in the residue properties such as hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility used in SVM★. It is thus of interest to examine whether SVM★ is useful for classification of homologous enzymes of different functions.

In this work, SVM★ is tested on four pairs of homologous enzymes of different families. These enzyme pairs are shown in Table VIII. Mixed results are obtained. While all eight enzymes 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 enzymes studied here. It is however difficult to accurately assess the capability of SVM★ classification of homologous enzymes of different functions based on the small number of homologous enzymes studied here. Further analysis is needed to provide a more objective assessment.

A Limitation of the SVM Classification Systems Developed in this Work

The SVM classification systems developed in this work are based on the two-class classification platform. One class contains proteins in a particular enzyme family, and another class consists of representative proteins outside of this family that includes both enzymes of the remaining 45 enzyme families and non-enzymes. For those enzymes that are simultaneously classified into more than one enzyme families, our classification systems may not be able to uniquely predict which family each of these enzymes belongs to.

Of the 8,291 enzymes correctly classified in this work, 6,658 or 80.3% of them are uniquely classified into a specific enzyme family using a scoring function.[30] Overall, the majority of the enzymes can be uniquely predicted by our classification systems, suggesting that our classification systems have certain level of unique prediction capability. None-the-less, the capability of unique prediction needs to be further enhanced by introducing methods that can further classify the non-uniquely classified enzymes into specific enzyme family. Multi-class classification approach[27] may be employed for such a purpose. In the multi-class classification approach, 46 additional SVM enzyme classification systems are trained, each from a positive set of all the enzymes in each enzyme family and a negative set of all the enzymes in the remaining 45 enzyme families. The non-uniquely classified enzymes are then tested against the 46 additional SVM classification systems. The unique family for each of these enzymes might be predicted either as that with the largest decision function value or by pair-wise classification with respect to multiple families.[27] Work is in progress to use a more comprehensive set of enzymes as the training sets and to develop the multi-class SVM enzyme classification systems by using more than 80,000 distinct enzyme sequence entries found from protein sequence databases.

CONCLUDING REMARKS

Our study suggests the potential of SVM in classification of enzymes into functional families. Moreover, it shows certain level of capability for classification of distantly related enzymes and homologous enzymes of different functions. When classifying an unknown protein, one does not know which family it might belong to. A screening process can be designed to scan all the families to determine which family it belongs to. Such a screening approach is also useful for classification of proteins that belong to multiple families. Further improvements on protein functional family coverage, sample collection, multi-class prediction models, and classification algorithm may enable the development of SVM into a useful tool for facilitating protein function prediction. Effort is being made to use a more comprehensive set of enzymes as the training sets to train SVM classification systems and to incorporate multi-class SVM classification systems to further enhance the unique prediction accuracy of our systems.

REFERENCES

1. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genomes and back. J Mol Biol 1998;283:707–725.
2. Eisenberg D, Marcotte CA, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000;405:823–826.
3. Huynen M, Snel B, Lathe W, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 2000;10:1204–1210.
4. Teichmann SA, Mitchison G. Computing protein function. Nat Biotechnol 2000;18:27–27.
5. Pellegrini M. Computational methods for protein function analysis. Curr Opin Chem Biol 2001;5:46–50.
6. Baxevanis AD. Practical aspects of multiple sequence alignment. Methods Biochem Anal 1998;39:172–188.
7. Bork P, Koonin EV. Predicting functions from protein sequences—where are the bottlenecks? Nat Genet 1998;18:313–318.
8. Schuler GD. Sequence alignment and database searching. Methods Biochem Anal 1998;39:145–171.
9. Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 1998;8:163–167.
10. Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, Knecht L. Functional inferences from reconstructed evolutionary biology involving rectified databases—an evolutionarily grounded approach to functional genomics. Res Microbiol 2000;151:97–106.
11. Fujiwara Y, Asogawa M. Protein function prediction using hidden Markov models and neural networks. NEC Res Dev 2002;43:238–241.
12. Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol 2001;11:354–363.
13. Di Gennaro JA, Siew N, Hoffman BT et al. Enhanced functional annotation of protein sequences via the use of structural descriptors. J Struct Biol 2001;134:232–245.
14. Enright AJ, Iliopoulos I, Kyrpides N, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86–90.
15. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. Science 1999;285:751–753.
16. Aravind L. Guilt by association: contextual information in genome analysis. Genome Res 2000;10:1074–1077.
17. Bock JR, Gough DA. Predicting protein-protein interactions from primary structure. *Bioinformatics* 2001;17:455–462.
18. Hodges HC, Tsai JW. 3D-Motifs: An informatics approach to protein function prediction. FASB J 2002;16:A543–A543.
19. Jensen LJ, Gupta R, Bolm N et al. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 2002;319:1257–1265.
20. Enright AJ, Ozounis CA. GeneRage: a robust algorithm for sequence clustering and domain detection. Bioinformatics 2000;16: 451–457.
21. Enright AJ, Van Dongen SV, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30:1575–1584.
22. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L. Gene families: the taxonomy of protein paralogs and chimeras. Science 1997;278:609–614.
23. Ponting CP. Issues in predicting protein function from sequence. Brief Bioinform 2001;2:19–29.

24. Burges CJC. A tutorial on Support Vector Machine for pattern recognition. Data Min Knowl Disc 1998;2:121–167.
25. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002;18: 147–159.
26. Cai CZ, Wang WL, Chen YZ. Support vector machine classification of physical and biological datasets. Inter J Mod Phys C 2003; 14:575–585.
27. Ding CHQ, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 2001;17:349–358.
28. Yu CS, Wang JY, Yang JM et al. Fine-grained protein fold assignment by support vector machines using generalized *n*peptide coding schemes and jury voting from multiple-parameter sets. Proteins 2003;50:531–536.
29. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48:566–570.
30. Hua SJ, Sun ZR. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 2001;308:397–407.
31. Cai YD, Liu XJ, Xu XB, Chou KC. Prediction of protein structural classes by support vector machines. Comput Chem 2002;26:293–296.
32. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 2002;30:47–49.
33. Enzyme nomenclature committee. Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. San Diego: Academic Press; 1992.
34. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31: 365–370.
35. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
36. Dubchak I, Muchnik I, Holbrook SR, Kim S-H. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci USA 92:8700–8704.
37. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 2003;31: 3692–3697.
38. Cai YD, Liu XJ, Xu XB, Chou KC. Support vector machines for predicting HIV protease cleavage sites in protein. J Comput Chem 2002;23:267–274.
39. Cortes C, Vapnik V. Support vector networks. Machine Learning 1995;20:273–297.
40. Cristianini N, Shawe-Taylor J. An introduction to support vector machines. United Kingdom: Cambridge University Press; 2001.
41. Baldi P, Brunak S, Chauvin Y, Anderson CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16**:**412–419.
42. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
43. Frishman D, Argos P. Recognition of distantly related protein sequences using conserved motifs and neural networks. J Mol Biol 1992;228:951–962.
44. Miyata Y, Nishida E. Distantly related cousins of MAP kinase: biochemical properties and possible physiological functions. Biochem Biophys Res Commun 1999;266:291–295.
45. Nagano N, Porter CT, Thornton JM. The (betaalpha)(8) glycosidases: sequence and structure analyses suggest distant evolutionary relationships. Protein Eng 2001;14:845–855.
46. Abagyan R, Frishman D, Argos P. Recognition of distantly related proteins through energy calculations. Proteins 1994;19**:**132–140.
47. Yang AS. Structure-dependent sequence alignment for remotely related proteins. Bioinformatics 2002;18:1658–1665.
48. Ishiguro M, Suzuki M, Takio K, Matsuzawa T, Titani K. Complete amino acid sequence of rat liver cytosolic alanine aminotransferase. Biochemistry 1991;30:6048–6053.
49. Hirai S, Izawa M, Osada S, Spyrou G, Ohno S. Activation of the JNK pathway by distantly related protein kinases, MEKK and MUK. Oncogene 1996;12:641–650.
50. Turner AJ, Hyde RJ, Lim J, Hooper NM. Structural studies of aminopeptidase P: a novel cellular peptidase. Adv Exp Med Biol 1997;421:7–16.
51. Lacy DB, Tepp W, Cohen AC, DasGupta BR, Stevens RC. Crystal structure of botulinum neurotoxin type A and implications for toxicity. Nat Struct Biol 1998;5:898–902.
52. Lowther WT, Matthews BW. Structure and function of the methionine aminopeptidases. Biochim Biophys Acta 2000;1477:157–167.
53. Bull HB, Breese K. Surface tension of amino acid solutions: a hydrophobicity scale of amino acid residue. Arch Biochem Biophys 1974;161:665–670.
54. Lin T-Y, Timasheff SN. On the role of surface tension in the stabilization of globular proteins. Protein Sci 1996;5:372–381.
55. Todd AE, Orengo CA, Thornton JM. Plasticity of enzyme active sites. Trends Biochem Sci 2002;27:419–426.
56. Todd AE, Orengo, CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.