THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 281, NO. 9, pp. 5357–5363, March 3, 2006

Prediction of Mitochondrial Proteins Using Support Vector Machine and Hidden Markov Model*[S]

Received for publication, October 11, 2005, and in revised form, December 7, 2005 Published, JBC Papers in Press, December 8, 2005, DOI 10.1074/jbc.M511061200

Manish Kumar, Ruchi Verma, and Gajendra P. S. Raghava[1]

From the Institute of Microbial Technology, Sector 39-A, Chandigarh 160036, India

Mitochondria are considered as one of the core organelles of eukaryotic cells hence prediction of mitochondrial proteins is one of the major challenges in the field of genome annotation. This study describes a method, MitPred, developed for predicting mitochondrial proteins with high accuracy. The data set used in this study was obtained from Guda, C., Fahy, E. & Subramaniam, S. (2004) *Bioinformatics* 20, 1785–1794. First support vector machine-based modules/methods were developed using amino acid and dipeptide composition of proteins and achieved accuracy of 78.37 and 79.38%, respectively. The accuracy of prediction further improved to 83.74% when split amino acid composition (25 N-terminal, 25 C-terminal, and remaining residues) of proteins was used. Then BLAST search and support vector machine-based method were combined to get 88.22% accuracy. Finally we developed a hybrid approach that combined hidden Markov model profiles of domains (exclusively found in mitochondrial proteins) and the support vector machine-based method. We were able to predict mitochondrial protein with 100% specificity at a 56.36% sensitivity rate and with 80.50% specificity at 98.95% sensitivity. The method estimated 9.01, 6.35, 4.84, 3.95, and 4.25% of proteins as mitochondrial in *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, mouse, and human proteomes, respectively. MitPred was developed on the above hybrid approach.

The mitochondrion, popularly known as the power house of the cell, is the central unit of eukaryotic cells.[2] It is a double membrane-bounded organelle with two spaces, the outer intermembrane space and inner matrix. Due to the presence of an explicit mitochondrial genome, unlike other organelles, its function is regulated by two genomes. It performs a plethora of biochemical reactions like oxidative phosphorylation, Krebs cycle, β-oxidation of fatty acids, DNA replication, transcription, translation, etc., some of which occur in mitochondria only. In addition mitochondria are also involved in apoptosis (1) and ionic homeostasis (2). Because of their multidimensional utility, mitochondrial proteins are associated with several human diseases, including Alzheimer disease (3), Type II diabetes (4) and Parkinson disease (5).

A majority of mitochondrial proteins are synthesized in cytoplasm from where they are transported inside mitochondria. But a small number of mitochondria-resident proteins are also synthesized inside mitochondria by the mitochondrial genome. Proteins that are imported to mitochondria contain a leader sequence at the N terminus that contains all the information needed to localize to mitochondria (6). But this is not true for all mitochondrial proteins. In many cases the leader sequence is altogether absent. This poses a major challenge in predicting mitochondrial proteins *in silico.*

In the past, a number of methods have been developed to predict the mitochondrial proteins, although most were not intended exclusively for mitochondrial proteins. Existing prediction methods can be divided into four categories. The similarity search-based techniques fall under the first category in which the query sequence is searched against experimentally annotated proteins. If the query protein has significant sequence similarity with any mitochondrial protein then it is predicted as a mitochondrial protein. But this method fails to predict new/novel proteins if these proteins do not have similarity with known proteins. Although the similarity-based method is very informative and considered to be best, it becomes severely handicapped when no apparent homology is found (7). In the second category, the methods are based on predicting signal sequences in proteins. A number of methods fall under the second category where sorting signals, present on the protein itself, are used for prediction. This category includes TargetP (8), SignalP (9), and PSORT II (10). Although these methods are quite popular, their major limitation is that not all proteins have signals; for example, only around 25% of yeast mitochondrial proteins have "matrix-targeting signals" particularly at the N terminus (11). Because of this, these methods fail to predict the proteins that do not have a signal. In the third category, methods attempt to predict subcellular localization on the basis of sequence composition. Some popular methods in this category are ESLpred (12), HSLpred (13), NNPSL (7), and LOCSVMPSI (14). Although their overall performance is very good, accuracy of prediction of mitochondrial proteins is much lower than for proteins in other locations. It shows that mitochondrial protein localization is much more complex than seen otherwise. Hence prediction of mitochondrial proteins warrants special attention. Recently Guda *et al.* (15) developed a method, MITOPRED, which falls under the fourth category of prediction methods. This method is exclusively developed for predicting mitochondrial proteins with a maximum accuracy of 92.3% (Mathews' correlation coefficient (MCC),[3] 0.638). Its output is the combined score of two scoring methods that assign an arbitrary score on the basis of existence of Pfam domains, differences in the amino acid composition, and the pI value of the protein.

In the present study we tried to improve the prediction accuracy of mitochondrial proteins. First we carried out systematic analysis of amino acid composition of both mitochondrial and non-mitochondrial proteins, and then on the basis of the conclusion drawn we developed the prediction method. In our study we used a powerful machine learn-

* This work was supported by the Council of Scientific and Industrial Research and the Department of Biotechnology, Government of India. This report has Institute of Microbial Technology communication number 052/2005. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Figs. S1 and S2.

[1] To whom correspondence should be addressed: Bioinformatics Centre, Inst. of Microbial Technology, Sector 39-A, Chandigarh 160036, India. Tel.: 91-172-2690557 or -2690225; Fax: 91-172-2690632 or -2690585; E-mail: raghava@imtech.res.in.

[2] MitPred was developed on a hybrid approach to predict mitochondrial proteins, and additional data are available upon request.

[3] The abbreviations used are: MCC, Mathews' correlation coefficient; SVM, support vector machine; HMM, hidden Markov model; SAAC, split amino acid composition.

ing technique, support vector machine (SVM), for classifying proteins instead of the pI score used in MITOPRED.

MATERIALS AND METHODS

Data Sets and Evaluation—We used the same data set that was constructed to develop MITOPRED by Guda *et. al* (15). This data set contain 1432 mitochondrial and 8940 non-mitochondrial sequences. First both mitochondrial and non-mitochondrial proteins were randomly divided into five parts. Each of these five sets consists of one-fifth of mitochondrial (~ 287) and one-fifth of non-mitochondrial (~ 1788) proteins. For training and for testing and evaluating our methods, we used a 5-fold cross-validation technique in which four sets were used for training and the remaining set was used for testing. This process was repeated five times so that each set was used once for testing (16).

Amino Acid and Dipeptide Composition—The aim of calculating the composition of proteins is to transform the variable length of protein sequences to fixed length feature vectors. This is an important and most crucial step during classification of proteins using machine learning techniques because they require fixed length patterns. In addition the conversion of a protein sequence to a vector of 20 dimensions using amino acid composition will encapsulate the properties of the protein into the vector. In addition to amino acid composition, dipeptide composition was also used for classification that gave a fixed pattern length of 400. The advantage of dipeptide composition over amino acid composition is that it encapsulates information about the fraction of amino acids as well as their local order. The amino acid as well as dipeptide composition was calculated as described by Garg *et al.* (13). Both compositions were used as input to classify mitochondrial and non-mitochondrial proteins using SVM.

Split Amino Acid Composition (SAAC)—In the case of SAAC, variable length protein sequences were represented by a fixed length pattern of 60 instead of 20 in the case of standard amino acid composition. In SAAC each protein is divided into three parts: (i) 25 amino acids of the N terminus, (ii) 25 amino acids of the C terminus, and (iii) the region between these two. The rationale behind using this is the fact that percent composition of the whole sequence does not give proper weight to compositional bias, which is known to be present in mitochondrial protein termini. Hence the advantage of SSAC over standard amino acid composition is that it provides greater weight to proteins that have a signal at either the N or C terminus.

N-terminal Signal—It is a well reported fact that mitochondrial proteins contain signal sequence at their N termini, and this is used in TargetP for prediction (8). In this work we adopted a different strategy to model N-terminal signal. It was represented by a binary vector of 525 (25×21) representing the N-terminal 25 residues. This binary vector was used to classify the proteins using SVM.

Combination of SVM-based Method and BLAST Search—BLAST (17) is the most commonly used tool for similarity search. In this study we computed the performance of BLAST in detecting mitochondrial proteins using default parameters at an *e* value of 0.1. In this, proteins of the test sets were used as the query against a data base containing the proteins of corresponding training sets. The performance of BLAST was calculated in terms of correct hits for mitochondrial proteins. In the next step the search result was combined with the SVM-based method. Although BLAST remains the final referee, SVM predictions were used only for proteins where either BLAST did not find any significant hit or the most significant hit was a non-mitochondrial protein and vice versa. We also repeated the same procedure at different cutoff *e* values. If a protein has a hit with *e* value less than the cutoff then the BLAST prediction was used; otherwise the SVM-based prediction was considered.

Occurrence of Pfam Domains Using Hidden Markov Models (HMMs)—HMMs are a class of probabilistic models that are generally applicable to time series or linear sequences. It describes probability distribution over a potentially infinite number of sequences. Basically it is a more advanced and sophisticated version of the Markov chain. In this study, HMM was implemented using HMMER (hmmer.wustl.edu/). The HMM-based sequence search was done using the Pfam data base (release 17.0) (18), which is a data base of multiple sequence alignment of proteins belonging to the same family. It contains 7868 families, each representing a Pfam domain. Each mitochondrial and non-mitochondrial protein in our data set was searched against the Pfam data base using the HMM search at an *e* value threshold of $1e-5$. Search results were analyzed to detect three type of domains: (i) exclusively mitochondrial domains occurring only in mitochondrial proteins, (ii) exclusively non-mitochondrial domains occurring only in non-mitochondrial proteins, and (iii) shared domains occurring in both type of proteins. A protein was assigned as a mitochondrial protein if it contains even one exclusive mitochondrial domain.

Hybrid Approach—In the hybrid approach SVM and the HMM search were combined to exploit the benefits of both *de novo* prediction by SVM and the highly sophisticated HMM-based similarity search technique in a well annotated and curated domain data base, Pfam. Because domains are structural, functional, and evolutional units of proteins, it is well known that all proteins are made up of either single or multiple domains. Hence searching domain(s) can be a major step toward determining the localization of a protein that may give a new insight on probable function of a protein. But because all the domains are not characterized yet, instances where not even a single hit is found can be a real challenge. In these cases incorporation of SVM-based prediction can be a good alternative. In this way the hybrid approach should be a better way of prediction. First proteins in the test set were searched against a data base of exclusive mitochondrial and non-mitochondrial domains. A protein was assigned as a mitochondrial protein if it has an exclusive mitochondrial domain and was assigned as a non-mitochondrial protein if it has an exclusive non-mitochondrial domain. But if the protein does not have any exclusive domain then the SVM-based method was used for prediction.

Evaluation on Independent Data Set—Cross-validation is the most popular method to evaluate performance of a prediction method. But it has been shown in the past that performance of *n*-fold cross-validation is not completely unbiased (19). To assess the unbiased performance of any method one needs to evaluate it on an independent data set. Keeping this in mind we evaluated the performance of our method on an independent data set generated from "OrganelleDB," which is a curated data base of mitochondrial proteins (20). It contains 723, 412, 352, 320, and 99 mitochondrial proteins of yeast, *Drosophila*, *Caenorhabditis elegans*, human, and mouse, respectively.

Annotation of Proteomes—Five complete eukaryotic proteomes were downloaded from the European Bioinformatics Institute (www.ebi.ac.uk/integr8) representing three non-vertebrates (*S. cerevisiae*, *C. elegans*, and *Drosophila melanogaster*) and two vertebrates (human and mouse). On these proteomes, by using the hybrid approach of the Pfam search and SVM, an estimation of the total number of mitochondrial proteins was carried out.

Support Vector Machine—In this study we implemented SVM by using the SVM^{light} package (21), which allows us to choose a number of parameters and kernels (*e.g.* linear, polynomial, radial basis function, and sigmoid) or any user-defined kernel. Assuming that we have a number of patterns $X_i \in R^d$ $(i = 1, 2, \ldots . n)$ with corresponding target values y_i ∈{target value}. Here the target value is either +1 (representing a mitochondrial protein) or −1 (for non-mitochondrial proteins). SVM

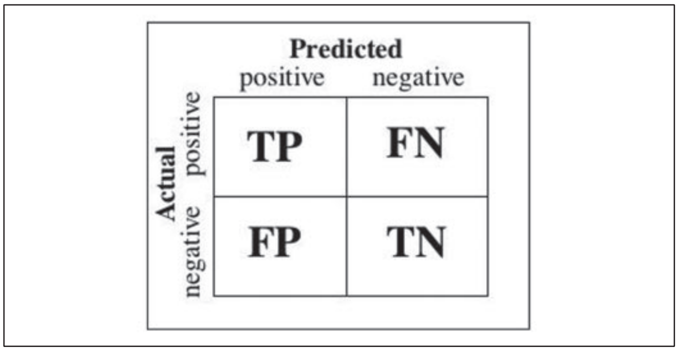

FIGURE 1. **Criteria of classification of a prediction into true positive (*TP*), true negative (*TN*), false positive (*FP*), or false negative (*FN*).** If a positive example is predicted as positive then it is classified under true positive prediction and vice versa for true negative prediction. But if a positive example is predicted as a member of a negative class and vice versa then it is classified as a false negative and false positive prediction, respectively.

maps the input vectors x_i into higher dimensional space with minimum error on the training set. The decision function implemented by SVM can be written as Equation 1.

$$F(x) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i K(x_i x_j + b)\right) \qquad \text{(Eq. 1)}$$

The value of α_i is given by the task of quadratic programming, thus maximizing the subject to $0 \leq \alpha_i \leq C$. C is the regulatory parameter that controls the trade-off between the margin and the training error, and b is the threshold for defining the hyperplane. The selection of kernel is very important in SVM, analogous to choosing architecture in artificial neural network. In this study, learning was carried out using three kernels: linear, polynomial, and sigmoid.

Evaluation Parameters—To assess the performance of methods we used several parameters routinely used in these types of studies (22 and 13). The following is a brief description of these parameters. (i) The sensitivity or percent coverage of mitochondrial proteins is the percentage of mitochondrial proteins correctly predicted as mitochondrial proteins. (ii) The specificity or percent coverage of non-mitochondrial proteins is the percentage of non-mitochondrial proteins correctly predicted as non-mitochondrial proteins. (iii) The accuracy is the percentage of correctly predicted proteins. These parameters can be calculated using Equations 2–4,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \qquad \text{(Eq. 2)}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \qquad \text{(Eq. 3)}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \qquad \text{(Eq. 4)}$$

where TP and TN are truly or correctly predicted positive (mitochondrial) and negative (non-mitochondrial) proteins, respectively (Fig. 1). FP and FN are falsely or wrongly predicted mitochondrial and non-mitochondrial proteins, respectively.

MCC is considered to be the most robust parameter of any class prediction method. An MCC equal to 1 is regarded as a perfect prediction, whereas 0 is for a completely random prediction.

TABLE 1

Result of BLAST search on data set used for MitPred

Percent coverage indicates the proteins that were predicted as mitochondrial proteins from the BLAST search. Correct hit shows proteins whose topmost hit belongs to the mitochondrial protein class. No hit is the number of proteins that did not get any hit below the threshold *e* value.

Data set	Number of mitochondrial proteins	Summary of BLAST hits			Percent coverage
		No hit	Total hits	Correct hit	
					%
Test1	286	31	255	235	82.16
Test2	287	105	182	111	38.67
Test3	287	66	221	174	60.62
Test4	286	42	244	206	72.03
Test5	286	86	200	164	57.34
Total/average	1432	330	1102	890	62.15

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \qquad \text{(Eq. 5)}$$

All the measures described above have a common drawback that they give the performance at a given threshold. A known threshold-independent parameter is receiver operating characteristic, which is a plot between true positive proportion (TP/TP + FN) and false positive proportion (FP/FP + TN). The area under the curve gave a single value to evaluate the performance of a method.

RESULTS AND DISCUSSION

Sequence Similarity Search—One of the common practices for predicting the function of a new protein is to perform a similarity search against a data base of well annotated proteins. In this study we used BLAST for predicting mitochondrial proteins using 5-fold cross-validation where four sets of mitochondrial and non-mitochondrial proteins were used to create a BLAST data base, and mitochondrial proteins of the corresponding test set were searched against this BLAST data base. This process was repeated five times so the BLAST search was performed once for each mitochondrial protein. As shown in Table 1, the performance of BLAST at default threshold varies from 38.67 to 82.16% with an average of 62.15%. This demonstrates that BLAST alone cannot predict all mitochondrial proteins.

SVM Modules of Amino Acid and Dipeptide Composition—It has been shown in the past that amino acid composition can be use to classify proteins. We analyzed amino acid composition of mitochondrial and non-mitochondrial proteins (Fig. 2). It was observed that amino acid composition of mitochondrial proteins was significantly different from that of non-mitochondrial proteins. Thus it is possible to discriminate mitochondrial proteins from other proteins. Thus an SVM-based classifier was developed using amino acid composition as input vector of dimension 20. Different kernels and parameters of SVM were tried and achieved accuracy of 78.37% with an MCC of 0.43 using radial basis function kernel where sensitivity and specificity is nearly the same (Table 2). It has also been observed in the past that dipeptide composition-based methods are more successful than amino acid composition-based methods in classification of proteins (23). Thus, an SVM-based module was developed to predict mitochondrial proteins using dipeptide composition as input vector of dimension 400. But contrary to our expectation, the performance of the dipeptide composition-based method was only marginally better than the amino acid composition-based method (Table 2). Although the receiver operating characteristic plot was also found to be far above the base line of random prediction (Fig. 3), still there are chances of improvement. We also

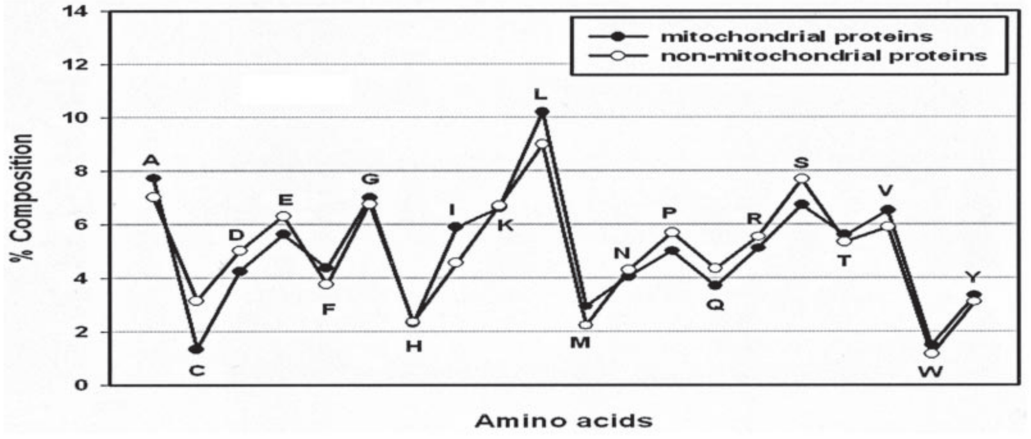

FIGURE 2. **Average percent composition of each of 20 amino acids in mitochondrial and non-mitochondrial proteins.**

TABLE 2

Performance of SVM with different inputs

Amino acid composition, amino acid composition used as input; Dipep. comp., dipeptide composition used as input; NT-25, amino acid composition of the N-terminal 25 amino acids used as input; CT-25, amino acid composition of the C-terminal 25 amino acids used as input; NT-25+R, the whole protein was divided into two parts, the N-terminal 25 amino acids and the remaining sequence, and the amino acid composition of both fragments was determined and together used as input (vector of 40 dimensions); Split, the whole protein was divided into three parts, the N-terminal 25 amino acids, the C-terminal 25 amino acids, and the remaining sequence, and the amino acid composition of all three fragments was determined and together used as input (vector of 60 dimensions); AUC, area under the curve. *J*, *c*, *t*, and *d* are the parameters used during training of SVM. Values in parentheses represent the maximum accuracy and MCC achieved with those particular parameters of SVM.

Input parameters	Parameter	Threshold	Sensitivity	Specificity	Accuracy	MCC	AUC
			%	%	%		
Amino acid composition	$J = 5, g = 0.001, c = 100$	−0.3	78.49	78.36	78.37 (88.23)	0.43 (0.46)	0.85
Dipep. comp.	$J = 5, t = 1, d = 3$	−0.3	77.03	79.75	79.38 (88.09)	0.44 (0.47)	0.86
NT-25	$J = 8, t = 1, d = 4$	−0.2	73.38	73.06	73.10 (88.69)	0.34 (0.44)	0.82
CT-25	$J = 6, d = 4,$	−0.2	64.40	64.33	63.48 (86.22)	0.19 (0.20)	0.69
NT-25+R	$J = 5, t = 2, g = 0.0001, c = 75$	−0.4	82.40	82.45	82.44 (90.426)	0.51 (0.56)	0.89
Split	$J = 3, g = 0.0001, c = 10$	−0.4	82.75	82.74	83.74 (90.39)	0.52 (0.57)	0.90

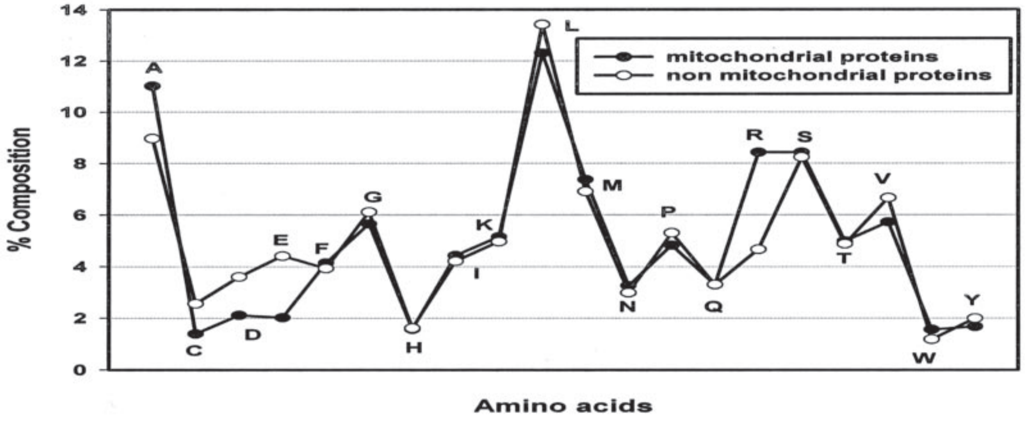

FIGURE 3. **Average percent composition of the first 15 N-terminal amino acid residues of mitochondrial and non-mitochondrial proteins.**

computed area under the curve for SVM modules and achieved areas under the curve of 0.85 and 0.86 for amino acid and dipeptide composition, respectively.

Combination of BLAST Search and SVM Module—Both the similarity search-based method (BLAST) and the machine learning method have their strengths and weaknesses. Similarity-based methods are better than any other method if the query sequence has significant similarity with any experimentally annotated protein. But it fails in the absence of similarity. On the other hand the machine learning method SVM is a general method in which the performance does not depend on similarity between target and query sequences. To improve the performance both the BLAST search and SVM were combined. In this, the training set was taken as the data base, and the proteins from test set were used as query proteins to search against this data base. Each protein from the test set was used to query against the corresponding data base. Among all the hits only the top hit (minimum *e* value) was considered as significant. If the hit corresponded to a mitochondrial protein then the query protein was directly assigned as a mitochondrial protein. Similarly if the hit corresponded to a non-mitochondrial protein then the query protein was directly assigned as a non-mitochondrial protein. Although the overall performance of the similarity search was quite good, it was observed that there was no hit for a large number of proteins. For these

TABLE 3

Performance of the combined approach of BLAST and SVM

Here the SVM module trained on split amino acid composition (N-terminal 25 residues, C-terminal 25 residues, and remaining amino acids) was used in combination with BLAST.

e value threshold	Sensitivity	Specificity	Accuracy	MCC
	%	%	%	
1*e*−1	77.950	91.404	89.266	0.638
1*e*−2	78.298	90.838	88.846	0.632
1*e*−3	78.648	90.508	88.622	0.626
1*e*−4	78.856	89.996	88.224	0.620
1*e*−5	78.856	89.996	88.224	0.620
1*e*−10	78.720	89.058	87.416	0.606
1*e*−20	78.438	87.714	83.238	0.580
1*e*−30	78.158	86.870	85.486	0.564
1*e*−40	78.438	86.488	85.208	0.560

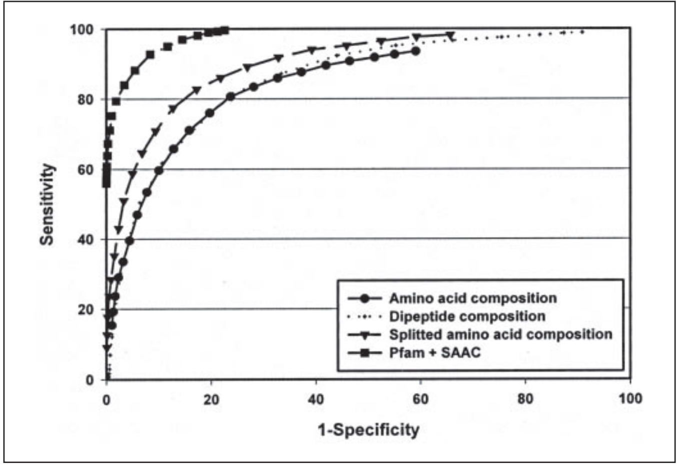

FIGURE 4. **Performance of different (amino acid composition, dipeptide composition, and split amino acid composition) SVM modules and the hybrid approach of SVM and an HMM search in a threshold-independent manner by receiver operating characteristic plot.**

proteins SVM was used to predict whether it is a mitochondrial or non-mitochondrial protein. In this study different *e* values from BLAST were used to assign location. The maximum accuracy using this strategy was around 89% with MCC around 0.63 (Table 3), which is better than the performance of the BLAST search or SVM module alone. This demonstrates that the combination of the similarity search and machine learning techniques can achieve very high accuracy if used to supplement each other (24).

Analysis of N-terminal and C-terminal Residues—In previous studies it has been shown that around 25% of proteins have an N-terminal signal; thus it is pertinent to analyze the amino acid composition of mitochondrial and non-mitochondrial proteins. Recently Guda *et al.* (15) analyzed the 25 residues of the N terminus and the remaining protein (after removing the 25 residues of the N terminus) for proteins of different cellular locations. They observed a strong bias in amino acid composition among them. It was also found that the composition of N-terminal residues and remaining residues of mitochondrial proteins was different. We extended their study to analyze the amino acid composition of mitochondrial proteins as follows: (i) 15, 20, 25, 30, and 35 N-terminal residues (supplemental Fig. S2a); (ii) remaining residues of the protein (after removing the 15, 20, 25, 30, and 35 N-terminal residues) (supplemental Fig. S2b); (iii) 15, 20, 25, 30, and 35 C-terminal residues (supplemental Fig. S2c); and (iv) remaining residues of the protein (after removing the 15, 20, 25, 30, and 35 C-terminal residues) (supplemental Fig. S2d). It was observed that the composition of N-terminal residues from 15 to 35 shows the same trend (Fig. 4). This was also the case when composition was determined without taking into consideration the N-terminal 15–35 residues. When these two compositions were compared, it was found that there is a clear difference between the composition at the N terminus of the protein and the remaining part of the protein. Although residues at the C terminus have a different composition then the rest of the protein, it is not as significant as for the N-terminal residues. We performed similar amino acid composition analysis for non-mitochondrial proteins. We did not find any significant difference between amino acid composition of N-terminal, C-terminal, and remaining residues of non-mitochondrial proteins. We also compared the composition of mitochondrial and non-mitochondrial proteins. As shown in supplemental Fig. S1, N-terminal residues of mitochondrial proteins are quite different from those of non-mitochondrial proteins.

SVM Module Based on N-terminal Sequences—Based on the above observations, we developed a method using N-terminal residues. Sequences were represented by binary matrix; for example for 25 N-terminal residues, a matrix of 25 × 21 was used in which a residue at a given position will be represented by vector of 21 (23). The SVM modules based on 15, 20, 25, 30, and 35 N-terminal residues were developed and achieved a maximum MCC from 0.28 to 0.31.

SVM Modules Based on Composition of N-terminal, C-terminal, and Remaining Protein Residues—It was observed that the amino acid composition of the N-terminal, C-terminal, and remaining mitochondrial protein residues was different from that of non-mitochondrial proteins. Thus we developed SVM modules based on the following: (i) composition of N-terminal residues with input vector 20, (ii) N-terminal residues and remaining residues with input vector 40; and (iii) N-terminal residues, C-terminal residues, and remaining residues using input vector 60. The maximum MCC of SVM modules based on composition of 15, 20, 25, 30, and 35 varies from 0.4 to 0.5, which is much better than the SVM module based on N-terminal sequence. This was surprising to us as the N-terminal sequence is supposed to contain complete information, whereas the composition has the total number of residues without order information. It seems that order of residues is not important for mitochondrial signals, but their presence is important; this is opposite to the known biological fact. This may be because all mitochondrial proteins do not have a leader signal, which is ultimately restricting the SVM to map the relationship between the N-terminal sequence and their localization. The performance of SVM modules with N-terminal residue composition and remaining residue composition (input vector 40) varies from an MCC of 0.5 to 0.6 that is better than SVM modules based on N-terminal residues or full protein. Because the performance of SVM modules using 15–35 C-terminal residues was very poor (MCC ~ 0.20), we dropped it from further analysis. The performance of SVM module-based split amino acid composition (N-terminal, C-terminal, and remaining residues) called SAAC, where input vector was 60, had an MCC from 0.5 to 0.6.

Hybrid Method: Pfam Domain and SVM Modules—In the method developed by Guda *et al.* (15) prediction of mitochondrial proteins was done on the basis of occurrence of Pfam domains. We adopted a similar strategy in this study (see "Materials and Methods"). All proteins in our data set were searched using HMMER against the Pfam data base and a total of 1662 domains was found. Among 1662 domains 206 were found exclusively in mitochondrial, 1162 were found exclusively in non-mitochondrial, and 147 were found in both type of proteins. A data base called MitoPfam, was built that consists of all three types of domains. To predict whether a protein can be localized in mitochondria or not, we performed an HMM search against the MitoPfam data base. A protein

TABLE 4

Combined result of Pfam search and SVM

Threshold is for cutoff for SVM on the basis of which performance is calculated. Bold is for the threshold where sensitivity and specificity are roughly equal.

Threshold	Sensitivity	Specificity	Accuracy	MCC
	%	%	%	
−2.0	99.580	77.408	80.934	0.596
−1.8	99.230	78.770	82.022	0.610
−1.6	98.950	80.498	83.432	0.628
−1.4	98.114	82.542	85.02	0.648
−1.2	96.926	85.488	87.308	0.676
−1.0	94.972	88.314	89.374	0.708
−0.8	**92.738**	**91.626**	**91.804**	**0.748**
−0.6	88.202	94.522	93.518	0.778
−0.4	84.014	96.554	94.560	0.800
−0.2	79.404	98.138	95.160	0.812
0.0	75.216	98.970	95.194	0.812
0.2	71.098	99.34	94.850	0.796
0.4	67.326	99.684	94.540	0.782
0.6	63.838	99.764	94.052	0.762
0.8	60.834	99.894	93.684	0.746
1.0	59.440	99.932	93.494	0.74
1.2	57.830	99.958	93.264	0.73
1.4	56.922	99.972	93.128	0.722
1.6	56.362	100.000	93.062	0.722
1.8	56.152	100.000	93.028	0.718
2.0	55.942	100.000	92.996	0.716

was assigned as mitochondrial if it has an exclusive mitochondrial domain or as non-mitochondrial protein if it has an exclusive non-mitochondrial domain. Using this approach 798 mitochondrial and 5732 non-mitochondrial proteins were assigned. It was found that there was no hit for a large number of proteins either due to the absence of an exclusive mitochondrial or non-mitochondrial domain or to no domain present at all. Thus we developed a hybrid method that combines the SVM module based on split amino acid composition and the occurrence of Pfam domain. In the hybrid method a protein is predicted to be mitochondrial or non-mitochondrial if it has an exclusive domain. In the case where a protein does not have any exclusive domain the SVM module based on SAAC was used for prediction. The performance of this hybrid method was evaluated at various thresholds of SVM module ranging from −2 to +2 (Table 4).

Performance on Independent Data Set—In our previous study (19), we observed a bias in the performance of a method on data used for testing and training despite jackknife testing (*e.g.* 5-fold cross-validation). Thus it is important to evaluate a newly developed method on an independent data set for unbiased evaluation. The independent data set used in this study consists of 723 yeast, 412 *Drosophila*, 352 *C. elegans*, 320 human, and 99 mouse mitochondrial proteins obtained from OrganelleDB (20). Our method predicted 571, 361, 198, 277, and 76 proteins corresponding to yeast, *Drosophila*, *C. elegans*, human, and mouse, respectively, of this data set as mitochondrial at default parameters. When the same sequences were submitted for prediction on the MITOPRED server at default parameters (confidence cutoff, 85%), for yeast, *Drosophila*, *C. elegans*, human, and mouse, 480, 304, 160, 249, and 71, proteins, respectively, were predicted as mitochondrial proteins (Table 5). It has been demonstrated by Guda *et al.* (15) that MITOPRED is better than existing methods like PSORT and TargetP. In their study they have clearly shown that MITOPRED shows better performance than other prediction methods like PSORT and TargetP. But the performance on the independent data set clearly shows that our method has performed even better than MITOPRED despite the fact that both are developed on the same data set. The possible reason behind this may be the fact that we have used both N- and C-terminal composition of proteins along with the composition of the remaining protein (SAAC), which is clearly a better parameter than terminal amino acid composition as shown by the performance of SVM modules.

TABLE 5

Performance of MitPred and MITOPRED servers on mitochondrial proteins retrieved from OrganelleDB

The prediction was done at default parameters of the web server.

Organism	Total number of proteins	Number of proteins predicted as mitochondrial	
		MitPred	MITOPRED
Yeast	723	571	480
C. elegans	352	198	160
Drosophila	412	361	304
Mouse	99	76	71
Human	320	277	249

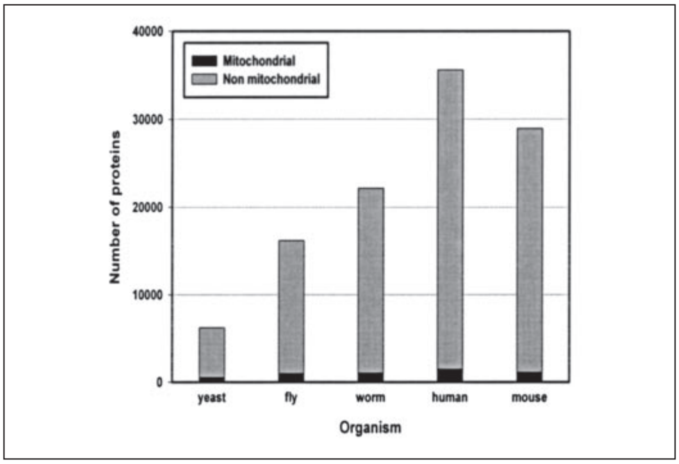

FIGURE 5. **Number of proteins whose potential subcellular location is predicted as mitochondrial in five representative proteomes by MitPred algorithm.**

Annotation of Proteomes—We chose six complete proteomes, ranging from single celled budding yeast to more complex *Drosophila*, *C. elegans*, human, and mouse, for annotation. The number of proteins predicted as mitochondrial is shown in Fig. 5. For yeast, MitPred predicted 561 proteins as mitochondrial among the total 6226, which is ∼9% of the total proteome. This is less than what others have estimated (like Guda *et al.* (15) and Marcotte *et al.* (11) who estimated 10% of the total proteomes or ∼750 proteins as determined experimentally (25)), but the difference is obviously due to the fact that we adopted the most stringent condition during annotation to filter out the false positives. For *Drosophila* and *C. elegans*, the number of mitochondrial proteins predicted was 1027 (6.3%) and 1071 (4.8%), respectively, from the complete proteome of 16,177 and 22,137 proteins. In *Drosophila* MITOPRED has estimated that ∼6.35% is mitochondrial protein; this tallies closely with that of our estimate. For *C. elegans* although Guda *et al.* (15) estimated 4%, Marcotte *et al.* (11) reported 4.3%. On the other hand for mouse and human 1144 and 1514 proteins were predicted as mitochondrial proteins from the complete proteome set of 28,936 and 35,595 proteins, respectively. In the case of human our method is consistent with the estimation of 1500 proteins by Taylor *et al.* (26), Lopez *et al.* (27), and Guda *et al.* (15). Recently Cameron *et al.* (28) have predicted the mitochondrial proteins in human by using a very novel and intelligent approach. Using MitoProt, they first predicted the proteins of yeast whose subcellular localization was likely to be mitochondria. Taking these proteins as the query, by using TBLASTN they identified the human proteins that are likely to be mitochondrial. In addition by using several stringent filters they predicted 361 human mitochondrial proteins that share close homology with yeast mitochondrial proteins. We propose that one of the main reasons behind the difference in the number of predicted mitochondrial proteins between the current method and MITOPRED is the difference between the numbers of proteins in the data downloaded from the European Bioinformatics Institute.

Web Server—The method presented here is available on the World Wide Web in the form of a server, "MitPred." The World Wide Web address is available upon request. The user can enter a protein sequence in any standard format such as FASTA. The server has the option to choose any of the following: SVM, BLAST + SVM, and Pfam search + SVM.

REFERENCES

1. Gottlieb, R. A. (2000) *Crit. Rev. Eukaryot. Gene Expr.* **10,** 231–239
2. Jassem, W., Fuggle, S. V., Rela, M., Koo, D. D. & Heaton, N. D. (2002) *Transplantation* **73,** 493–499
3. Hutchin, T. & Cortopassi, G. A. (1995) *Proc. Natl. Acad. Sci. U. S. A.* **92,** 6892–6895
4. Gerbitz, K. D., Gempel, K. & Brdiczka, D. (1996) *Diabetes* **45,** 113–126
5. Orth, M. & Schapira, A. H. (2002) *Neurochem. Int.* **40,** 533–541
6. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2000) *Molecular Biology of the Cell*, p. 678, 4th Ed., Garland Science, New York
7. Reinhardt, A. & Hubbard, T. (1998) *Nucleic Acids Res.* **26,** 2230–2236
8. Emanuelsson, O., Nielsen, H., Brunak, S. & Heijne, G. (2000) *J. Mol. Biol.* **300,** 1005–1016
9. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. & Miyano, S. (2002) *Bioinformatics* **18,** 298–305
10. Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnady, G. E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K. & Brinkman, F. S. (2003) *Nucleic Acids Res.* **31,** 3613–3617
11. Marcotte, E. M., Xenarios, I., van Der Bliek, A. M. & Eisenberg, D. (2000) *Proc. Natl. Acad. Sci. U. S. A.* **97,** 12115–12120
12. Bhasin, M. & Raghava, G. P. S. (2004) *Nucleic Acids Res.* **32,** W414–W419
13. Garg, A., Bhasin, M. & Raghava, G. P. S. (2005) *J. Biol. Chem.* **280,** 14427–14432
14. Xie, D., Li, A., Wang, M., Fan, Z. & Feng, H. (2005) *Nucleic Acids Res.* **33,** W105–110
15. Guda, C., Fahy, E. & Subramaniam, S. (2004) *Bioinformatics* **20,** 1785–1794
16. Kaur, H. & Raghava, G. P. S. (2003) *Protein Sci.* **12,** 627–634
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410
18. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). *Nucleic Acids Res.* **32,** D138–D141
19. Bhasin, M. & Raghava, G. P. S. (2004b) *Bioinformatics* **20,** 421–423
20. Wiwatwattana, N. & Kumar, A. (2005) *Nucleic Acids Res.* **33,** D598–D604
21. Joachims, T. (1999) in *Advances in Kernel Methods—Support Vector Learning* (Scholkopf, B., Burges, C. & Smola, A., eds) MIT Press, Cambridge, MA
22. Kumar, M., Bhasin, M., Natt, N. K. & Raghava, G. P. S. (2005) *Nucleic Acids Res.* **33,** 154–159
23. Bhasin, M. & Raghava, G. P. S. (2004) *J. Biol. Chem.* **279,** 23262–23266
24. Bhasin, M., Garg, A. & Raghava, G. P. S. (2005) *Bioinformatics* **21,** 2522–2524
25. Sickmann, A., Reinders, J., Wagner, Y., Joppich, C., Zahedi, R., Meyer, H. E., Schonfisch, B., Perschil, I., Chacinska, A., Guiard, B., Rehling, P., Pfanner, N. & Meisinger, C. (2003) *Proc. Natl. Acad. Sci. U. S. A.* **100,** 13207–13212
26. Taylor, S. W., Fahy, E. & Ghosh, S. S. (2003) *Trends Biotechnol.* **21,** 82–88
27. Lopez, M. F., Kristal, B. S., Chernokalskaya, E., Lazarev, A., Shestopalov, A. I., Bogdanova, A. & Robinson, M. (2000) *Electrophoresis* **21,** 3427–3440
28. Cameron, J. M., Hurd, T. & Robinson, B. H. (2005) *Bioinformatics* **21,** 1825–1830