balt3/zbc-bc/zbc-bc/zbc01205/zbc9396-05g eichelbt S=4 9/2/05 20:57 Comments: ARTNO: M411789200

THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 280, No. ??, Issue of ???? ??, pp. 1–xxx, 2005

Printed in U.S.A.

AQ: A

Support Vector Machine-based Method for Subcellular Localization of Human Proteins Using Amino Acid Compositions, Their Order, and Similarity Search*[S]

Received for publication, October 18, 2004, and in revised form, January 12, 2005
Published, JBC Papers in Press, January 12, 2005, DOI 10.1074/jbc.M411789200

Aarti Garg, Manoj Bhasin, and Gajendra P. S. Raghava‡

From the Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India

Here we report a systematic approach for predicting subcellular localization (cytoplasm, mitochondrial, nuclear, and plasma membrane) of human proteins. First, support vector machine (SVM)-based modules for predicting subcellular localization using traditional amino acid and dipeptide (i + 1) composition achieved overall accuracy of 76.6 and 77.8%, respectively. PSI-BLAST, when carried out using a similarity-based search against a nonredundant data base of experimentally annotated proteins, yielded 73.3% accuracy. To gain further insight, a hybrid module (hybrid1) was developed based on amino acid composition, dipeptide composition, and similarity information and attained better accuracy of 84.9%. In addition, SVM modules based on a different higher order dipeptide i.e. i + 2, i + 3, and i + 4 were also constructed for the prediction of subcellular localization of human proteins, and overall accuracy of 79.7, 77.5, and 77.1% was accomplished, respectively. Furthermore, another SVM module hybrid2 was developed using traditional dipeptide (i + 1) and higher order dipeptide (i + 2, i + 3, and i + 4) compositions, which gave an overall accuracy of 81.3%. We also developed SVM module hybrid3 based on amino acid composition, traditional and higher order dipeptide compositions, and PSI-BLAST output and achieved an overall accuracy of 84.4%. A Web server HSLPred (www.imtech.res.in/raghava/hslpred/ or bioinformatics.uams.edu/raghava/hslpred/) has been designed to predict subcellular localization of human proteins using the above approaches.

AQ: B

The successful completion of a human genome project has yielded huge amount of sequence data. Analysis of this data to extract the biological information can have profound implications on biomedical research. Therefore, mining of biological information or functional annotation of piled up sequence data is a major challenge to the modern scientific community. Determination of functions of all of these proteins using experimental approaches is a difficult and time-consuming task. Traditionally, the similarity search-based tools has been used for functional annotations of proteins (1). This approach fails when unknown query protein does not have significant homology to proteins of known functions. The functions of the proteins are closely related to its cellular attributes, such as subcellular localization and its association with the lipid bilayer (subcellular localization) (2, 3); hence, the related proteins must be localized in the same cellular compartment to cooperate toward a common function (4). In addition, information on the localization of proteins with known function may provide insight about its involvement in specific metabolic pathways (5–7). Therefore, an attempt has been made to predict subcellular localization of proteins to elucidate the function.

Several methods have been devised earlier to predict the subcellular localization of the eukaryotic and prokaryotic proteins using different approaches and data sets (8). The most commonly used approach utilizes alignment or similarity search against an experimentally annotated data base. But this approach fails in the absence of significant similarity between the query and target protein sequences (1). Another popular approach is based on identification of sequence motifs such as signal peptide or nuclear localization signal (9). This approach has been limited by the observation that all of the proteins residing in a compartment do not have universal motif. To overcome these limitations, several machine learning technique-based methods, such as artificial neural networks and support vector machines (SVMs),[1] have been developed to predict the subcellular localization of proteins. These methods are based on the several features of protein sequences such as recognition of N-terminal sorting signals or the composition of amino acids. These methods predict subcellular localization either for prokaryotic or eukaryotic proteins, such as PSORT (10) and TargetP (11) for eukaryotes and SubLoc (8) and NNPSL (1) for both prokaryotes and eukaryotes, with good accuracy (>70%). Recently, our group has also developed a new hybrid approach-based method, ESLPred, which predicts the four major subcellular localizations (nuclear, cytoplasmic, mitochondrial, and extracellular) of eukaryotic proteins with an overall accuracy of 88% (12). To the best of our knowledge, there is no method for the prediction of subcellular localization of human proteins. Availability of sequence data of human genes in recent years demands a reliable and accurate method for prediction of subcellular localization of human proteins.

AQ: C

Fn1

In the present study, a systematic attempt has been made to develop a method for the subcellular localization of human proteins. The SVM modules based on different features of the proteins such as amino acid composition and dipeptide composition of proteins have been constructed. In addition, a similarity search-based module, HuPSI-BLAST, has also been developed, using PSI-BLAST to predict the localization of human

* This work was supported by the Council of Scientific and Industrial Research and the Department of Biotechnology, Government of India. This report has IMTECH communication number 050/2004. 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 supplementary information, including nine additional tables.

‡ To whom correspondence should be addressed. Tel.: 91-172-2690557; Fax: 91-172-2690632; E-mail: raghava@imtech.res.in.

AQ: L

[1] The abbreviations used are: SVM, support vector machine; MARS, multivariate adaptive regression splines; RI, reliability index.

balt3/zbc-bc/zbc-bc/zbc01205/zbc9396-05g eichelbt S=4 9/2/05 20:57 | Comments: | ARTNO: M411789200

TABLE I
Number of sequences within each subcellular location group

Subcellular location	Number of sequences
Cytoplasm	840
Mitochondria	315
Nuclear	858
Plasma Membrane	1519
Endoplasmic Reticulum	63
Extracellular	48
Peroxisome	25
Lysosome	51
Golgi	32
Centrosome	8
Microsome	21
Total	3780

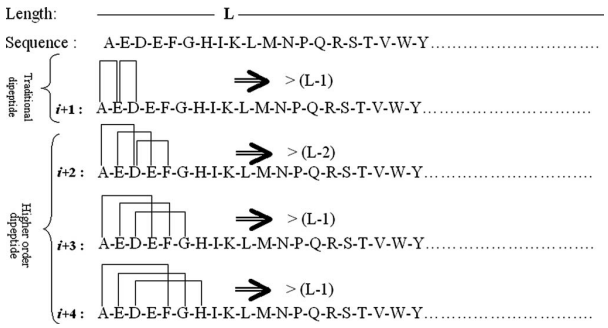

FIG. 1. **Graphic representation of traditional and higher order dipeptide compositions.**

proteins. Further, SVM module "hybrid1" has been developed using amino acid composition, traditional dipeptide composition, and results of PSI-BLAST prediction. The SVM modules based on higher order dipeptide compositions (i + 2, i + 3, and i + 4) and combinations of various feature-based modules have also been constructed. Here we have also compared the performance of the present organism-specific method (HSLPred) with ESLPred (12), a general method for prediction of subcellular localization of eukaryotic proteins. In addition, the performance of HSLPred has also been assessed on various mammalian and nonmammalian genomes and on an independent data set. It was observed that this method can predict the subcellular localization of human proteins and proteins from related genomes with high accuracy. In other words, our method can also be used for the prediction of subcellular localization of mammalian proteins.

MATERIALS AND METHODS

The Data Set—The data set of human proteins with experimentally annotated subcellular localization has been derived from release 44.1 of the SWISSPROT data base (13). Of 10,777 human proteins available in the data base, subcellular localization information was available for 7910 sequences. These 7910 sequences were screened strictly in order to develop a high quality data set for predicting subcellular localization of human proteins. The sequences annotated as "fragments," "isoforms," "potential," "by similarity," or "probable" were filtered out from the data set. Further, sequences residing in more than one subcellular location (such as a protein sequence labeled with "nuclear and cytoplasmic" or "mitochondrial and cytoplasmic") were also excluded from the data set. The sequence redundancy of data set was further reduced by using PROSET software (14) such that no two sequences had >90% sequence identity in the data set. The final data set consists of 3780 protein sequences that belong to 11 subcellular locations as shown in Table I. The number of sequences for the last seven subcellular locations was not sufficient for developing a prediction method. Therefore, a method was developed for only four major subcellular locations of human proteins (840 cytoplasmic, 315 mitochondrial, 858 nuclear, and 1519 plasma membrane).

Support Vector Machines—An excellent machine learning technique support vector machine has been used for the prediction of subcellular localization of human proteins. Previously, SVM has been successfully used for the classification of microarray data, MHC binder prediction, and protein secondary structure prediction (15, 16, 17). In the present study, a freely downloadable package of SVM, SVM_light has been used to predict the subcellular localization of proteins. The prediction of subcellular localization is a multiclass classification problem. Therefore, N SVMs for N class classification have been constructed. Here, the class number was equal to four for human proteins. The ith SVM was trained with all of the samples in the ith class with positive label and negative label for proteins of remaining subcellular localizations. This kind of SVM is known as one versus rest SVM (1-v-r SVM) (8). In this way, four SVMs were constructed for the subcellular localization of human proteins. An unknown sample was classified into the class that corresponds to the SVM with highest output score. We have adopted different approaches based on different features of a protein, such as amino acid composition and dipeptide composition, in the fixed length format.

Amino Acid Composition—Amino acid composition is the fraction of each amino acid in a protein. This representation completely misses the order of amino acids. The fraction of all 20 natural amino acids was calculated using Equation 1,

$$\text{Fraction of amino acid i} = \frac{\text{Total number of amino acid i}}{\text{Total number of amino acids in protein}} \quad \text{(Eq. 1)}$$

where i can be any amino acid.

Traditional Dipeptide Composition (i + 1)—Dipeptide composition was used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400 (20 × 20). This representation encompassed the information of the amino acid composition along with the local order of amino acids. The fraction of each dipeptide was calculated using Equation 2.

$$\text{Fraction of dep (i + 1)} = \frac{\text{Total number of dep (i + 1)}}{\text{Total number of all possible dipeptides}} \quad \text{(Eq. 2)}$$

where dep (i + 1) is one of 400 dipeptides. In addition, to observe the interaction of the ith residue with the 3rd, 4th, and 5th residue in the sequence, higher order dipeptides such as i + 2, i + 3, and i + 4, respectively (Fig. 1), were generated using Equation 3,

$$\text{Fraction of (i + n) dep} = \frac{\text{Total number of (i + n) dep}}{\text{Total number of all possible dipeptides}} \quad \text{(Eq. 3)}$$

where n = 2, 3, or 4, and dep (i + n) is one of 400 dipeptides.

Multivariate Adaptive Regression Splines—In this study, we also made an attempt to use a simple and reliable machine learning technique, multivariate adaptive regression splines (MARS), for predicting subcellular localization of human proteins. It has been shown previously that MARS performs as well as other machine learning techniques such as neural networks. In addition, MARS also provides information about the relative importance of different input variables for the classifications or predictions (18, 19). In the present study, we have downloaded the XTAL regression software package that incorporates the Xmars version of MARS for the SUN workstation (available on the World Wide Web at www.ece.umn.edu/groups/ece8591/xtal.html). This version of MARS uses a maximum of 10 predictable input variables. Thus, we have used compositions of amino acid properties for the prediction rather then amino acid and dipeptide compositions.

Compositions of Amino Acid Properties—We have used five commonly used properties of amino acids: (i) nonpolar aliphatic amino acids (Gly, Ala, Val, Leu, Ile, and Pro), (ii) polar uncharged amino acids (Ser, Thr, Cys, Met, Asn, and Gln), (iii) aromatic amino acids (Phe, Tyr, and Trp), (iv) negatively charged amino acids (Asp and Glu), and (v) positively charged amino acids (Lys, Arg, and His). In order to obtain the composition of a property, we added compositions of its residues (e.g. compositions of negatively charged residues would be compositions of Asp and Glu).

HuPSI-BLAST—A module HuPSI-BLAST was designed to predict subcellular localization of human proteins, in which the query sequence was searched against data base of human proteins using PSI-BLAST. The data base consists of 3532 sequences belonging to four major subcellular locations (cytoplasmic, mitochondrial, nuclear, and plasma

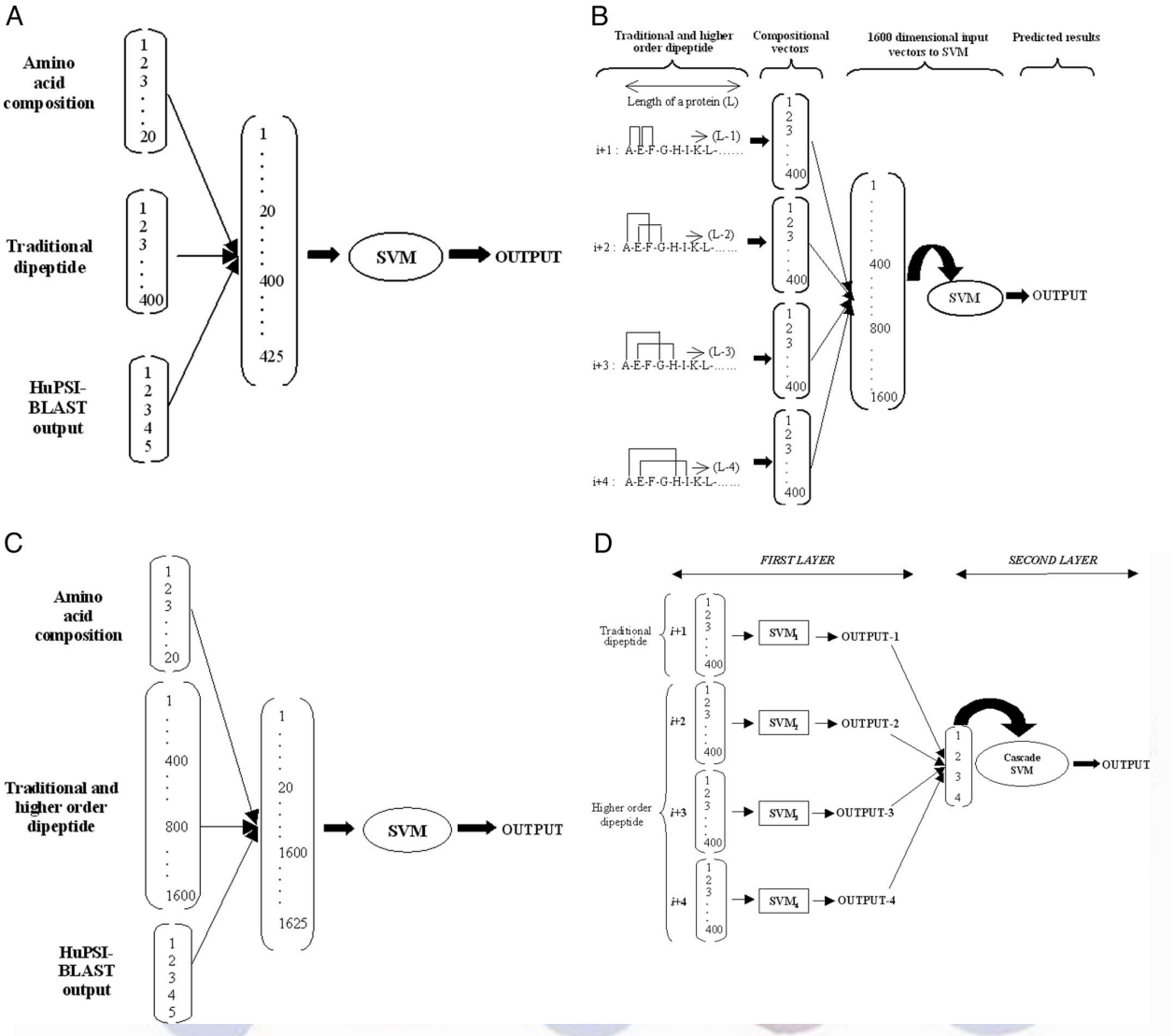

FIG. 2. a, the hybrid1 SVM module incorporates the features of a protein (amino acid and traditional dipeptide composition) and output of the HuPSI-BLAST module. b, the hybrid2 SVM module constructed using normal and higher order dipeptide compositions. c, the hybrid3 SVM module developed using a vector of 20 dimensions of amino acid composition, 1600 for traditional and higher order dipeptide compositions, and 5 of HuPSI-BLAST output. d, the SVM cascade consists of two layers of SVM.

membrane). The subcellular localization of these proteins has been proven experimentally. The PSI-BLAST was used instead of the normal standard BLAST to search the data base, because it has the capability to detect remote homologies (20). It carries out an iterative search in which the sequences found in one round of search are used to build a score model for the next round of searching. Three iterations of PSI-BLAST were carried out at a cut-off E value of 0.001. This module could predict any of the four localizations (cytoplasmic, mitochondrial, nuclear, or plasma membrane), depending upon the similarity of the query protein to the proteins present in the data base. The module would return "unknown subcellular localization" if no significant similarity was obtained.

AQ: F

Hybrid SVM Modules—Recently, our group has introduced the concept of a hybrid SVM module for the prediction of subcellular localization of eukaryotic proteins (12). In the present study, an attempt has been made to elaborate the concept of hybrid modules by designing hybrid modules based on different approaches. The description of the approaches used to develop different hybrid modules is described below.

Hybrid1 SVM Module—The hybrid1 SVM module encapsulates the information of amino acid composition, traditional dipeptide composition, and PSI-BLAST output (Fig. 2a). SVM was provided with an input vector of 425 dimensions that consisted of 20 for amino acid composition, 400 for dipeptide composition, and 5 for PSI-BLAST output. The PSI-BLAST output was converted to binary variables using the representation shown in Equation 4.

F2

$$\begin{pmatrix} \text{Cytoplasmic} & \rightarrow & 1\,0\,0\,0\,0 \\ \text{Mitochondrial} & \rightarrow & 0\,1\,0\,0\,0 \\ \text{Nuclear} & \rightarrow & 0\,0\,1\,0\,0 \\ \text{Plasma membrane} & \rightarrow & 0\,0\,0\,1\,0 \\ \text{Unknown} & \rightarrow & 0\,0\,0\,0\,1 \end{pmatrix} \quad \text{(Eq. 4)}$$

Hybrid2 SVM Module—The hybrid2 SVM module was constructed using all higher order dipeptide compositions (i + 2, i + 3, and i + 4) along with traditional dipeptide composition (i + 1). This hybrid2 module was provided with an input vector of 1600 dimensions, 400 from each dipeptide composition (Fig. 2b).

Hybrid3 SVM Module—The hybrid3 SVM module was constructed using amino acid composition, traditional dipeptide composition (i + 1), higher order dipeptide compositions (i + 2, i + 3, and i + 4), and similarity search-based results (Fig. 2c). The module was provided with input vector of 1625 dimensions, comprising 20 for amino acid compositions, 1600 for the above four types of dipeptide compositions, and 5 for PSI-BLAST output.

TABLE II
Detailed performance of various SVM modules developed using different features of a protein and PSI-BLAST

Approaches used	Cytoplasm		Mitochondria		Nuclear		Plasma membrane		Average	
	ACC[a]	MCC[b]	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Composition-based (A)	63.5	0.52	46.0	0.52	76.2	0.67	90.3	0.78	76.6	0.67
PSI-BLAST (B)	56.9		40.6		68.2		92.0		73.3	
Dipeptide-based i + 1 (C)	58.3	0.52	48.3	0.52	80.2	0.71	93.4	0.80	77.8	0.69
Hybrid1 (A + B + C)	75.4	0.67	69.8	0.68	82.4	0.79	94.8	0.89	84.9	0.80

[a] ACC, accuracy.
[b] MCC, Matthews correlation coefficient.

Evaluation of HSLPred—The performance of SVM modules constructed in this report was evaluated using a 5-fold cross-validation technique. In this technique, a relevant data set was partitioned randomly into five equal sized sets. The training and testing was carried out five times, each time using one distinct set for testing and the remaining four sets for training. To assess the predictive performance, the accuracy and Matthews correlation coefficient were calculated as described by Hua and Sun (8), using Equations 5 and 6,

$$\text{Accuracy}(x) = \frac{p(x)}{\text{Exp}(x)} \qquad \text{(Eq. 5)}$$

$$\text{MCC}(x) = \frac{p(x)n(x) - u(x)o(x)}{\sqrt{(p(x) + u(x))(p(x) + o(x))(n(x) + u(x))(n(x) + o(x))}} \qquad \text{(Eq. 6)}$$

where x can be any subcellular location (cytoplasmic, mitochondrial, nuclear, or plasma membrane), Exp(x) is the number of sequences observed in location x, p(x) is the number of correctly predicted sequences of location x, n(x) is the number of correctly predicted sequences not of location x, u(x) is the number of underpredicted sequences, and o(x) is the number of overpredicted sequences.

Reliability Index—The reliability index (RI) is a commonly used measure of prediction that provides confidence about the predictions to the users. The RI assignment is a useful indication of the level of certainty in the predictions for the particular sequence. The strategy used for assigning the RI is similar to that used in the past by our group (12). The RI was assigned according to the difference (Δ) between the highest and second highest SVM output scores. The reliability index for the hybrid1 approach-based module was calculated using Equation 7.

$$\text{RI} = \begin{cases} \text{INT}(\Delta * 5/3 + 1) & \text{if } 0 \le \Delta < 4, \\ 5 & \text{if } \Delta \ge 4\,. \end{cases} \qquad \text{(Eq. 7)}$$

In order to validate the performance of HSLPred and to compare with other methods such as ESLPred (12), two other data sets were also used. A brief description is as follows.

Independent Data Set—Techniques such as cross-validation and bootstrapping are routinely used for evaluating the performance of any method. Still, the best way of testing the performance of a newly developed method is to test it on an independent data set that contains the patterns used neither during training nor during testing of the method. Independent data were derived from the latest release, 45.2, of the SWISSPROT data base (13). This data set contained 164 human proteins (30 cytoplasmic, 11 mitochondrial, 60 nuclear, and 63 plasma membrane) and was not used in the training and testing of the HSLPred method.

ESLPred Data Set—To compare the performance of the present method (HSLPred) with ESLPred, another method developed by our group for subcellular localizations of eukaryotic proteins (12), the data set of ESLPred was used. ESLPred was trained on 2427 eukaryotic proteins (1097 nuclear, 684 cytoplasmic, 321 mitochondrial, and 325 extracellular). This data set was further divided into two main sets: (a) mammalian and (b) nonmammalian (eukaryotic proteins other then mammalian) proteins, to assess the performance of HSLPred on these two different systems.

In addition, the data sets of other mammalian genomes such as rat, rabbit, bovine, and sheep have also been downloaded from the latest release, 45.2, of the SWISSPROT data base (13), to check the generalizability of HSLPred on other closely related genomes. The data set used is shown in Table S6 of the Supplementary Material.

RESULTS AND DISCUSSION

Human genome sequencing has produced sequences of more than >40,000 genes. Amazingly, genes are simple, consisting of four types of nucleotides (adenine, guanine, cytosine, and thymine) and get translated into far more complex proteins that are made up of 20 different types of amino acids. The four types of nucleotides in various different orders carry information for making the specific proteins that directs the make up of each human being. Among many other things, proteins control human development and physiology and provide resistance to diseases. In order to perform its appropriate functions, each protein must be translocated to its correct intra- or extracellular compartments. Hence, subcellular localization is a key step characteristic of each functional protein. AQ: G

Since 1991, numerous algorithms have been developed to predict subcellular localization of proteins, based on amino acid compositions (21), neural network (1) covariant discriminant algorithm (22), Markov chains (23), and support vector machines (8, 24). Recently, Gardy et al. (25) have developed a tool, PSORT-B, that combines several methods for the prediction of subcellular localization for Gram-negative bacterial proteins. In general, artificial intelligence-based techniques such as SVM and artificial neural networks are considered as elegant approaches for the prediction of subcellular localization of proteins. AQ: H

The performance of all of the SVM modules developed in this study has been evaluated through a 5-fold cross-validation technique. The SVM training has been carried out by the optimization of various kernel function parameters and the value of the regularization parameter C. The detailed results obtained using various kernel function parameters have been shown in the Supplementary Material, Table S1. It has been observed that the RBF kernel performs better than linear and polynomial kernels in the case of the amino acid composition-based SVM module. Thus, for all of the SVM modules developed in the present study, the RBF kernel has been used.

The amino acid composition-based SVM module (kernel = RBF, $\gamma = 300$, C = 2, j = 1) has been able to achieve an overall accuracy of 76.6% for all of the four subcellular localizations (Table II). Further, to implement information about frequency as well as the local order of residues, an SVM module based on traditional dipeptide compositions has been constructed. The traditional dipeptide (i + 1) composition-based SVM module has achieved the best results (77.8%) with the RBF kernel ($\gamma = 50$, C = 6, j = 1). This accuracy is nearly 1% better than the amino acid composition-based SVM module. The detailed performance of amino acid- and traditional dipeptide composition-based SVM modules in assigning different subcellular localizations has been shown in Table II. T2

The homology of a protein with other related sequences provides a broad range of information about the protein. Hence, similarity search-based module HuPSI-BLAST has been constructed to encapsulate evolutionary information of the proteins. During 5-fold cross-validation, no significant hits have been obtained for 671 of 3532 proteins. Therefore, the performance of this module is poorer in comparison with amino acid composition- as well as dipeptide composition-based modules. This module has predicted cytoplasmic, mitochondrial, nuclear,

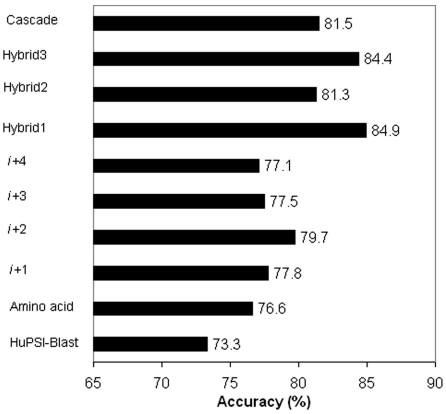

FIG. 3. **Comparison of overall performance of SVM modules constructed on the basis of different features and approaches.**

and plasma membrane subcellular localizations with 56.9, 40.6, 68.2, and 92% accuracy, respectively, and achieved an overall accuracy of 73.3% (Table II). It proves that compositions (amino acid and dipeptide) can annotate the data more reliably in comparison with the similarity search-based tool.

To further, enhance the prediction accuracy, methodologies such as "hybrids" have been devised to encapsulate more comprehensive information of the proteins. The first hybrid SVM-based module, hybrid1, has been constructed using amino acid composition, traditional dipeptide composition, and PSI-BLAST results. The hybrid1 module with the RBF kernel (γ = 50, C = 2, j = 1) has achieved striking overall accuracy of 84.9%, which is significantly better then rest of the modules developed in this study. These results confirm that prediction accuracy of subcellular localization of proteins can be increased using a wide range of information about a protein.

In addition, higher order dipeptide (i + 2, i + 3, and i + 4) composition-based SVM modules have been constructed to examine the effect of different positions of amino acids on the subcellular localization. The overall performance of higher order dipeptide compositions in predicting subcellular localization is shown in Fig. 3. The (i + 2) dipeptide composition-based SVM module has achieved an overall accuracy of 79.7%, ~2% higher in comparison with the traditional and other higher order dipeptide composition-based SVM modules. It has also been observed that an accuracy of i + 2 dipeptide composition-based modules is nearly 2% more for cytoplasmic proteins, and for the remaining three subcellular localizations (mitochondrial, nuclear, and plasma membrane), it is almost comparable with traditional dipeptide compositions. Further, the performance of i + 3 and i + 4 dipeptide composition-based modules has been found to be similar to the traditional dipeptide composition-based SVM module (Fig. 3).

Since the i + 2 dipeptide composition-based module has achieved better accuracy in comparison with traditional dipeptide composition, different hybrid modules have been constructed with an aim to increase the overall accuracy. The SVM module hybrid2 has been constructed using all higher order dipeptide compositions (i + 2, i + 3, and i + 4) along with traditional dipeptide compositions. The overall accuracy of the hybrid2 SVM module is 3% less than the hybrid1 module, but it is nearly 4% higher in comparison with traditional dipeptide composition. This proves that it is able to encapsulate more information, which is useful in delineating the proteins of different subcellular localizations. Furthermore, another SVM module hybrid3 has been constructed using amino acid compositions, traditional dipeptide compositions (i + 1), higher order dipeptide compositions (i + 2, i + 3, and i + 4), and PSI-BLAST results. However, the hybrid3 SVM module has been predicted with an overall accuracy of 84.4%, which is nearly equal to the hybrid1 module. Further enhancement in accuracy cannot be achieved due to the complexity of input patterns, since the hybrid3 module has been provided with an input vector of 1625 dimensions.

In addition to hybrid modules, a cascade SVM-based approach has also been adopted to classify the human proteins with better accuracy. The cascade SVM consists of two layers of SVM (Fig. 2d). The first layer consists of models based on traditional and higher order dipeptide compositions (i + 1, i + 2, i + 3, and i + 4), and the second layer consists of an SVM model that correlates the output of the first layer model and provides a final output. The cascade SVM module has been able to achieve an accuracy of 81.5%, comparable with the performance of the hybrid2 module. A comparison of the accuracies of all of the SVM modules developed on the basis of different approaches is shown in Fig. 3.

To evaluate the prediction reliability, RI assignment has been carried out for the hybrid1 SVM module. It indicates the effectiveness of an approach in the prediction of subcellular localization of proteins. The RI is a measure of confidence in the prediction. Ideally, the accuracy and probability of a correct prediction should increase with an increase of RI values. We have computed the average prediction accuracy of proteins having an RI value greater than or equal to n, where n = 1,2 . . . 5. As shown in Table S9 of the Supplementary Material, HSLPred has been able to predict 67.3% of sequences with an average prediction accuracy of 94.9% at RI $\geq$ 5. This demonstrates that a user can predict a large number of sequences with higher accuracy for RI $\geq$ 5. Similarly, HSLPred has been able to predict 83.4% sequences with an accuracy of 91.1% for RI $\geq$ 3.

The main objective of the present study was to develop a method for the subcellular localization of human proteins. Since the present method has been trained on the specific organism's proteins, it should be more accurate and better for the particular organism in comparison with methods such as ESLPred, developed generally for all eukaryotic proteins. The following analysis has been performed to show the superiority of HSLPred over existing methods such as ESLPred.

First, the performance of HSLPred has been evaluated on proteins used to develop the ESLPred method. The hybrid1-based approach of the HSLPred method has been able to predict cytoplasmic, mitochondrial, and nuclear proteins (of ESLPred) with an accuracy of 91.8, 35.2, and 78.3%, respectively, and an overall accuracy of 76.1% has been attained. The details have been given in the Supplementary Material, Table S3.

Second, in order to examine the performance of the ESLPred method on human proteins, we have applied the ESLPred method on proteins used to develop HSLPred. It has been observed that the hybrid-based approach of ESLPred predicted cytoplasmic, mitochondrial, and nuclear proteins with an accuracy of 42.7, 57.8, and 84.8%, respectively. An overall accuracy of 62.9% has been achieved. For details, see Table S4 of the Supplementary Material. These results indicated that the performance of an organism-specific HSLPred method is better than ESLPred for predicting human proteins.

Furthermore, in order to check the reason behind poor performance of HSLPred in comparison with ESLPred on eukaryotic proteins, the data set used to develop the ESLPred method has been divided into two main sets: (i) mammalian and (ii) nonmammalian (all eukaryotic proteins other than mammalian) proteins. These two sets have been further predicted using the HSLPred server. We found that the HSLPred method has achieved an overall accuracy of 85 and 70.8% for mamma-

lian and nonmammalian protein sets, respectively, as shown in the Supplementary Material, Table S5. It proves that HSLPred can predict mammalian proteins with good accuracy and non-mammalian proteins with fair accuracy.

Further, the performance of both HSLPred and ESLPred has been assessed on an independent data set to estimate the unbiased performance of a method. It has been observed that HSLPred has been able to predict 20, 7, 50, and 58 proteins correctly out of 30, 11, 60, and 63 (cytoplasmic, mitochondrial, nuclear, and plasma membrane proteins), respectively, using the hybrid1 module. An overall accuracy of 82.3% has been achieved, whereas the ESLPred method has been able to achieve an overall accuracy of 64.4%. The detailed results have been shown in Table S2 of the Supplementary Material. In summary, the performance of HSLPred has been found to be better during both cross-validation and testing of an independent data set, suggesting that it is not an artifact. We have also tested the generalizability of the HSLPred algorithm with other genomes such as rat, rabbit, bovine, and sheep to assess the predictive performance of HSLPred on other closely related genomes. It has been observed that HSLPred also predicts other mammalian proteins with considerably high accuracy. The detailed results obtained have been shown in Table S7 of the Supplementary Material. Hence, the HSLPred method can also be used for the prediction of subcellular localization of other closely related mammalian proteins. In other words, it can act as a generalized method for various closely related mammalian genomes.

Although SVM and artificial neural networks are powerful techniques for the classification of proteins, they have their own limitations, since these techniques produce results that are sometimes difficult to interpret. Since subcellular localization has resulted from a number of input variables including hydrophobicity, amino acid composition, homology to other localized proteins, and localization motifs, the interpretation of results can provide new insights into protein subcellular localization. In the present study, we also used the MARS technique (18, 19) for the classification of subcellular localization of human proteins using five given properties of amino acids. It has been observed that for the classification of cytoplasmic proteins, composition of negatively charged amino acids (Asp and Glu) plays an important role. However, for the classification of mitochondrial proteins, the relative importance of positively charged (Lys, Arg, and His) and polar uncharged (Ser, Thr, Cys, Met, Asn, and Gln) amino acids has been observed. In the case of nuclear proteins, composition of aromatic amino acids (Phe, Tyr, and Trp) and for the plasma membrane proteins composition of positively charged amino acids (Lys, Arg, and His) has been found to be important. The detailed results obtained have been shown in Table S8 of the Supplementary Material. Further, cytoplasmic, mitochondrial, nuclear, and plasma membrane proteins have achieved accuracy of 35.7, 21.9, 50.0, and 82.4%, respectively, and an overall accuracy of 58% has been attained. In order to account for this lower accuracy, either due to the use of MARS or the specified properties of input variables, we have developed an SVM module based on the inputs used for MARS. We observed that the accuracy achieved by the SVM module (60.8%) was slightly better than MARS, demonstrating that MARS is also a powerful technique for the classification of proteins. Here, we like to comment that performance of MARS can be further improved if amino acid or dipeptide compositions are used as input variables.

HSLPred Server—Various types of SVM modules constructed in the present study have been implemented on a Web server (HSLPred) using CGI/Perl script. The HSLPred server is available on the World Wide Web at www.imtech.res.in/raghava/hslpred/ or bioinformatics.uams.edu/raghava/hslpred/. Users can enter a protein sequence in one of the standard formats, such as FASTA, GenBank™, EMBL, GCG, or plain format. The server provides options to select various approaches for the prediction of the subcellular localization of a query sequence. In the case of the default prediction, it uses the hybrid1 module for prediction. An overall architecture of the HSLPred server is shown in Fig. 4. F4

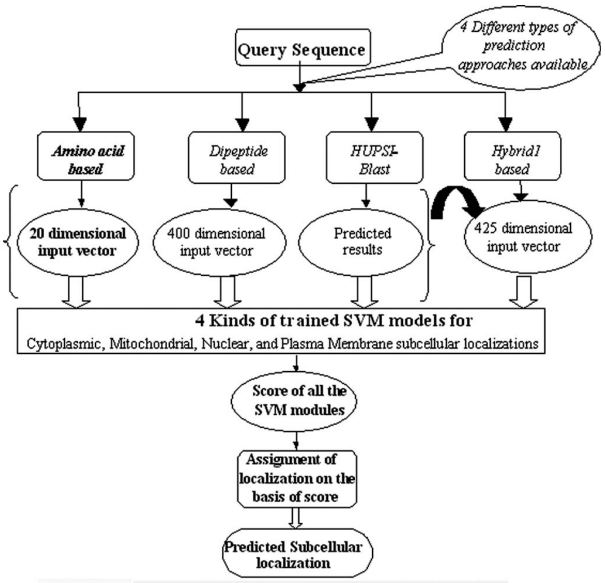

FIG. 4. **Overall architecture of the HSLPred server.**

Acknowledgments—We thank anonymous reviewers for excellent suggestions and Dr. D. Sarkar for correcting grammatical mistakes.

REFERENCES

1. Reinhardt, A., and Hubbard, T. (1998) Nucleic Acids Res. **26,** 2230–2236
2. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994) Molecular Biology of the Cell, 3rd Ed., Garland Publishing, New York AQ: I
3. Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P., and Darnell, J. (1995) Molecular Cell Biology, 3rd Ed., Scientific American Books, New York AQ: J
4. Nair, R., and Rost, B. (2003) Nucleic Acids Res. **13,** 3337–3340
5. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., and Herrmann, R. (1996) Nucleic Acids Res. **24,** 4420–4449
6. Nakai, K., and Kanehisa, M. (1991) Proteins **11,** 95–110
7. Nakai, K., and Kanehisa, M. (1992) Genomics **14,** 897–911
8. Hua, S., and Sun, Z. (2001) Bioinformatics **17,** 721–728
9. Fujiwara, Y., and Asogawa, M. (2001) Genome Inform. Ser. Workshop Genome Inform. **12,** 103–112
10. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyana, S. (2002) Bioinformatics **18,** 298–305
11. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) J. Mol. Biol. **300,** 1005–1016
12. Bhasin, M., and Raghava, G. P. S. (2004) Nucleic Acids Res. **32,** 415–419
13. Bairoch, A., and Apweiler, R. (2000) Nucleic Acids Res. **28,** 45–48
14. Brendel, V. (1991) Math. Comput. Modelling **16,** 37–43
15. Brown, M. P. S., Grundy, W. N., Lion, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., and Haussler, D. (2000) Proc. Natl. Acad. Sci. U. S. A. **97,** 262–297
16. Donnes, P., and Elofsson, A. (2002) BMC Bioinformatics **3,** 25 AQ: K
17. Ward, J. J., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2003) Biochemistry **19,** 1650–1655
18. Friedman, J. H. (1991) Ann. Stat. **19,** 1–141
19. Friedman, J. H., and Tukey, J. W. (1974) IEEE Trans. Comput. **23,** 881–890
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) J. Mol. Biol. **215,** 403–410
21. Chou, K. C. (1995) Proteins **21,** 319–344
22. Chou, K. C., and Elrod, D. (1999) Protein Eng. **12,** 107–118
23. Yuan, Z. (1999) FBBS Lett. **451,** 23–26
24. Chou, K. C., and Cai, Y. D. (2002) J. Biol. Chem. **277,** 45765–45769
25. Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnady, G. E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., and Brinkman, F. S. (2003) Nucleic Acids Res. **13,** 3613–3617