JBC Papers in Press. Published on January 12, 2005 as Manuscript M411789200

SVM -based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search

Aarti Garg, Manoj Bhasin and G.P.S. Raghava*
Institute of Microbial Technology
Sector – 39A, Chandigarh, India

Running title: SVM-based method for subcellular localization of human proteins

***Address for Correspondence**

Dr. G. P. S. Raghava, Scientist	**Email: raghava@imtech.res.in**
Bioinformatics Centre	**Web: http://www.imtech.res.in/raghava/**
Institute of Microbial Technology	**Phone: +91-172-2690557**
Sector 39A, Chandigarh, INDIA	**Fax: +91-172-2690632**

Summary

Here we report a systematic approach for predicting subcellular localization (cytoplasm, mitochondrial, nuclear and plasma membrane) of human proteins. Firstly, 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 similarity-based search against non-redundant database of experimentally annotated proteins yielded 73.3% accuracy. To gain further insight, 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 module based on 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 (http://www.imtech.res.in/raghava/hslpred/ or http://bioinformatics.uams.edu/raghava/hslpred/) has been designed to predict subcellular localization of human proteins using the above approaches.

Introduction

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 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 towards 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 datasets (8). The most commonly used approach utilizes alignment or similarity search against an experimentally annotated database. 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 (NLS) (9). This approach has been limited by the observation that all the proteins residing in a compartment do not have universal motif. To overcome these limitations, several machine learning techniques based methods such artificial neural networks (ANN) and support vector machines (SVM) 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, 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.

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 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 non-mammalian genomes and on an independent data set. It was observed that 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 dataset of human proteins with experimentally annotated subcellular localization has been derived from release 44.1 of the SWISSPROT database (13). Out of 10777 human proteins available in database, subcellular localization information was available for 7910 sequences. These 7910 sequences were screened strictly in order to develop a high-quality dataset for predicting subcellular localization of human proteins. The sequences annotated as "fragments", "isoforms", "potential", "by similarity", or "probable" were filtered out from the dataset. 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 dataset. The sequence redundancy of dataset was further reduced by using PROSET software (14) such that no two sequences had >90% sequence identity in the dataset. The final dataset consists 3780 protein sequences that belong to 11 subcellular locations as shown in the Table 1. The number of sequences for the last 7 subcellular locations was not sufficient for developing prediction method. Therefore, method was developed for only 4 major subcellular locations of human proteins (840 cytoplasmic, 315 mitochondrial, 858 nuclear, 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 microarry data, MHC binders 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 sub-cellular localization of proteins. The prediction of subcellular localization is a multi-class classification problem. So, N SVMs for N class classification have been constructed. Here, the class number was equal to four for human proteins. The i^{th} SVM was trained with all the samples in the i^{th} 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.

$$\text{Fraction of amino acid i} = \frac{\text{total number of amino acid i}}{\text{total number of amino acids in protein}} \quad \ldots\ldots 1$$

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 acid. The fraction of each dipeptide was calculated using equation 2.

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

Where, dep $(i+1)$ is one out of 400 dipeptide.

In addition, to observe the interaction of the i^{th} residue with the third, fourth and fifth residue in the sequence, higher order dipeptides such as *i*+2, *i*+3, and *i*+4 respectively (Figure 1) were generated using equation 3.

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

Where, n is equal to 2, 3 or 4; dep (*i*+n) is one out of 400 dipeptide.

Multivariate Adaptive Regression Splines

In this study, we also made an attempt to use simple and reliable machine learning technique Multivariate Adaptive Regression Splines (MARS) for predicting sub-cellular localization of human proteins. It has been shown previously that MARS performs as good 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 SUN workstation (http://www.ece.umn.edu/groups/ece8591/xtal.html). This version of MARS uses maximum 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) non-polar aliphatic amino acid (G, A, V, L, I, P), ii) polar uncharged amino acids (S, T, C, M, N, Q), iii) aromatic amino acids (F, Y, W), iv) negatively charged amino acids (D, E) and v) positively charged amino acids (K, R, H). In order to get composition of a property, we added compositions of its residues, for e.g. compositions of negatively charged residues would be compositions of D and E.

HuPSI-BLAST

A module HuPSI-BLAST was designed to predict subcellular localization of human proteins, in which the query sequence was searched against database of human proteins using PSI-BLAST. The database consists of 3532 sequences belonging to 4 major subcellular locations (cytoplasmic, mitochondrial, nuclear and plasma membrane). The subcellular localization of these proteins has been proven experimentally. The PSI-BLAST was used instead of normal standard BLAST to search the database 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 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 database. The module would return “unknown subcellular localization” if no significant similarity was obtained.

Hybrid SVM modules

Recently, our group has introduced the concept of 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 has been described below:

Hybrid1 SVM module

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

$$\left(\begin{array}{lcl} \text{Cytoplasmic} & \longrightarrow & 1\ 0\ 0\ 0\ 0 \\ \text{Mitochondrial} & \longrightarrow & 0\ 1\ 0\ 0\ 0 \\ \text{Nuclear} & \longrightarrow & 0\ 0\ 1\ 0\ 0 \\ \text{Plasma membrane} & \longrightarrow & 0\ 0\ 0\ 1\ 0 \\ \text{Unknown} & \longrightarrow & 0\ 0\ 0\ 0\ 1 \end{array}\right) \qquad \ldots\ldots4$$

Hybrid2 SVM module

The hybrid2 SVM module was constructed using all higher order dipeptide compositions (*i*+2, *i*+3, *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 compositions (Figure 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, *i*+4) and similarity search based results (Figure 2c). The module was provided with input vector of 1625 dimensions, comprising 20 for amino acid compositions, 1600 for above four types of dipeptide compositions and 5 for PSI-BLAST output.

Evaluation of HSLPred

The performance of SVM modules constructed in this report was evaluated using 5-fold cross-validation technique. In this technique, relevant dataset was partitioned randomly into 5 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, accuracy and Matthew’s correlation coefficient (MCC) were calculated as described by Hua and Sun, (8) using equation 5 and 6.

$$\text{Accuracy}(x) = \frac{p(x)}{Exp(x)} \qquad \ldots\ldots\ldots5$$

$$\text{MCC}(x) = \frac{p(x)n(x) - u(x)o(x)}{\sqrt{[p(x)+u(x)][p(x)+o(x)][n(x)+u(x)][n(x)+o(x)]}} \qquad \ldots\ldots\ldots\ldots6$$

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 under-predicted sequences and o(x) is the number of over-predicted 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 as 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 the equation 7.

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

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

Independent dataset

The 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 newly developed method is to test it on an independent dataset, that contains the patterns neither used during training and nor during testing of the method. An independent data was derived from the latest release 45.2 of the SWISSPROT database (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 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 dataset 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) non-mammalian (eukaryotic proteins other then mammalian) proteins, to assess the performance of HSLPred on these two different systems.

In addition, the data set of other mammalian genomes such as rat, rabbit, bovine and sheep have also been downloaded from the latest release 45.2 of the SWISSPROT database (13), to check the generalizability of HSLPred on other closely related genomes. The data set used is shown in Table S6 of supplementary material (http://www.imtech.res.in/raghava/hslpred/supl.html).

Results and Discussion

The human genome sequencing has produced sequences of more than >40,000 genes. Amazingly, genes are simple consisting of four type 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, physiology and provide resistance to diseases. In order to perform its appropriate functions, each protein must be translocated to its correct intra- or extra-cellular compartments. Hence, subcellular localization is a key step characteristic of each functional protein.

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 combined several methods together, for the prediction of subcellular localization for Gram-negative bacterial proteins. In general, artificial intelligence (AI) based techniques such as SVM and artificial neural networks are considered as elegant approaches for the prediction of subcellular localization of proteins.

The performance of all the SVM modules developed in this study has been evaluated through 5-fold cross-validation technique. The SVM training has been carried out by the optimization of various kernel function parameters and value of the regularization parameter C. The detailed results obtained using various kernel function parameters have been shown in the supplementary material Table S1 (http://www.imtech.res.in/raghava/hslpred/supl.html). It has been observed that RBF kernel performs better than linear and polynomial kernels, in the case of amino acid composition based SVM module. Thus, for all the SVM modules developed in the present study, RBF kernel has been used.

The amino acid composition based SVM module (kernel=RBF, γ=300, C=2, j=1) has been able to achieve an overall accuracy of 76.6% for all the 4 subcellular localizations (Table 2). Further, to implement information about frequency as well as local order of residues, 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 (γ=50, C=6, j=1). This accuracy is nearly 1% better than 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 2.

The homology of a protein with other related sequences provides 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 proteins out of 3532 proteins. Therefore, the performance of this module is poorer in comparison to amino acid composition as well as dipeptide composition based modules. This module has predicted cytoplasmic, mitochondrial, nuclear 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 2). It proves that compositions (amino acid and dipeptide) can annotate the data more reliably in comparison to similarity search based tool.

To further, enhance the prediction accuracy, the 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 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 wide range of information about a protein.

In addition, higher order dipeptide (*i*+2, *i*+3, and *i*+4) compositions 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 Figure 3. The (*i*+2) dipeptide composition based SVM module has achieved overall accuracy of 79.7%, ~2% higher in comparison to the traditional and other higher order dipeptide compositions based SVM modules. It has also been observed that accuracy of *i*+2 dipeptide composition based modules is nearly 2% more for cytoplasmic proteins and for remaining three subcellular localizations (mitochondrial, nuclear, plasma membrane), it is almost comparable to traditional dipeptide compositions. Further, the performance of *i*+3 and *i*+4 dipeptide composition based modules has been found to be similar to traditional dipeptide composition based SVM module (Figure 3).

Since, *i*+2 dipeptide composition based module has achieved better accuracy in comparison to traditional dipeptide composition, therefore, 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, *i*+4) along with traditional dipeptide compositions. The overall accuracy of hybrid2 SVM module is 3% lesser than hybrid1 module but it is nearly 4% higher in comparison to 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 modules hybrid3 has been constructed using amino acid compositions, traditional dipeptide compositions (*i*+1), higher order dipeptide compositions (*i*+2, *i*+3, *i*+4) and PSI-BLAST results. However, hybrid3 SVM module has been predicted with an overall accuracy of 84.4% that is nearly equal to hybrid1 module. Further enhancement in accuracy cannot be achieved due to complexity of input patterns as hybrid3 module has been provided with an input vector of 1625 dimensions.

In addition, to hybrid modules, 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 (Figure 2d). First layer consists of models based on traditional and higher order dipeptide compositions (*i*+1, *i*+2, *i*+3, *i*+4) and second layer consists of SVM model that correlates the output of first layer model and provides a final output. The cascade SVM module has been able to achieve an accuracy of 81.5%, comparable to the performance of hybrid2 module. The comparison of accuracies of all the SVM modules developed on the basis of different approaches is shown in Figure 3.

To evaluate the prediction reliability, RI assignment has been carried out for the hybrid1 SVM module. It tells about the effectiveness of an approach in the prediction of subcellular localization of proteins. The RI is a measure of confidence in the prediction. Ideally, accuracy and probability of correct prediction should increase with increase of RI values. We have computed the average prediction accuracy of proteins having RI value greater then equal to n where, n=1,2…5. As shown in Table S9 of supplementary material (http://www.imtech.res.in/raghava/hslpred/supl.html), HSLPred has been able to predict 67.3% of sequences with an average prediction accuracy of 94.9% at RI ≥ 5. This demonstrates that user can predict large number of sequences with higher accuracy for RI ≥ 5. Similarly, HSLPred has been able to predict 83.4% sequences with an accuracy of 91.1% for RI ≥ 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 to methods such as ESLPred, developed generally for all eukaryotic proteins. Following analysis has been performed to show superiority of HSLPred over existing methods such as ESLPred:

Firstly, the performance of HSLPred has been evaluated on proteins used to develop ESLPred method. The hybrid1-based approach of HSLPred method has been able to predict cytoplasmic, mitochondrial, and nuclear proteins (of ESLPred) with an accuracy of 91.8%, 35.2%, 78.3%, respectively, and an overall accuracy of 76.1% has been attained. The details have been given in the supplementary material Table S3 (http://www.imtech.res.in/raghava/hslpred/supl.html).

Secondly, in order to examine the performance of ESLPred method on human proteins, we have applied ESLPred method on proteins used to develop HSLPred. It has been observed that 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 supplementary material (http://www.imtech.res.in/raghava/hslpred/supl.html). These results indicated that the performance of organism specific method HSLPred is better than ESLPred for predicting human proteins.

Furthermore, in order to check the reason behind poor performance of HSLPred in comparison to ESLPred on eukaryotic proteins, the data set used to develop ESLPred method has been divided into two main sets: i) Mammalian and ii) Non-mammalian (all eukaryotic proteins other than mammalian) proteins. These two sets have been further predicted using HSLPred server. We found that HSLPred method has achieved an overall accuracy of 85% and 70.8% for mammalian and non-mammalian protein sets respectively, as shown in supplementary material Table S5 (http://www.imtech.res.in/raghava/hslpred/supl.html). 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, 58 proteins correctly out of 30, 11, 60, 63 (cytoplasmic, mitochondrial, nuclear and plasma membrane proteins) respectively, using hybrid1 module. An overall accuracy of 82.3% has been achieved. Where as, ESLPred method has been able to achieve overall accuracy of 64.4%. The detailed results have been shown in Table S2 of supplementary material (http://www.imtech.res.in/raghava/hslpred/supl.html). In summary, the performance of HSLPred has been found to be better both during cross-validation as well as 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 supplementary material (http://www.imtech.res.in/raghava/hslpred/supl.html). Hence, 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 ANN are powerful techniques for the classification of proteins, but they have their own limitations, as these techniques produce results, which are sometimes difficult to interpret. Since, subcellular localization has resulted from number of input variables including hydrophobicity, amino acid composition, homology to other localized proteins, localization motifs, hence, interpretation of results can provide new insights into protein subcellular localization. In the present study, we also used 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 (D and E) plays an important role. However, for the classification of mitochondrial proteins, the relative importance of positively charged (K, R and H) and polar uncharged (S, T, C, M, N, and Q) amino acids has been observed. In the case of nuclear proteins composition of aromatic amino acids (F, Y, and W) and for the plasma membrane proteins composition of positively charged amino acids (K, R and H) has been found to be important. The detailed results obtained have been shown in Table S8 of the supplementary material (http://www.imtech.res.in/raghava/hslpred/supl.html). 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 SVM module based on the inputs used for MARS. We observed that accuracy achieved by 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 as web server (HSLPred) using CGI/Perl script. HSLPred server is available at http://www.imtech.res.in/raghava/hslpred/ or http://bioinformatics.uams.edu/raghava/hslpred/. Users can enter 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 subcellular localization of a query sequence. In the case of default prediction, it uses the hybrid1 module for prediction. An overall architecture of HSLPred server has been shown in Figure 4.

Acknowledgements

The authors are thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial assistance. We are also grateful to anonymous reviewers for excellent suggestions and Dr. D. Sarkar for correcting the grammatical mistakes. This report has IMTECH communication no. 050/2004.

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., New York and London: Garland Publishing, Chapter 1
3. Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P., and Darnell, J. (1995) *Molecular cell biology,* 3rd Ed., New York: Scientific American Books; Chapter 3
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) *Mathl. Comput. Modelling* **16**, 37-43
15. Brown, M. P. S., Grundy, W. N., Lion, D., Cristianini, N., Sugnet, C. W., Furey, T. S., AresJr, M., and Haussler, D. (2000) *PNAS* 97, 262-97
16. Donnes, P., and Elofsson, A. (2002) BMC *Bioinformatics* **3**, 25
17. Ward, J. J., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2003) *Biochemistry* **19**, 1650-55
18. Friedman, J. H. (1991) *Ann. Stat.* **19**, 1-141
19. Friedman, J. H., and Tukey, J. W. (1974) *IEEE Trans. Computers*, **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 Engg.* **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

Figure Legends

Figure 1. The graphical representation of traditional and higher order dipeptide compositions
Figure 2a. The hybrid1 SVM module incorporates the features of a protein (amino acid and traditional dipeptide composition) and output of HuPSI-BLAST module.
Figure 2b. The hybrid2 SVM module constructed using normal and higher order dipeptide compositions
Figure 2c. 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.
Figure 2d. SVM cascade consists of two layers of SVM.
Figure 3. The comparison of an overall performance of SVM modules constructed on the basis of different features and approaches.
Figure 4. An overall architecture of HSLPred server.

Table 1. Number of sequences within each subcellular location groups

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

Table 2. 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	MCC	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**

ACC: Accuracy, MCC: Matthew's correlation coefficient

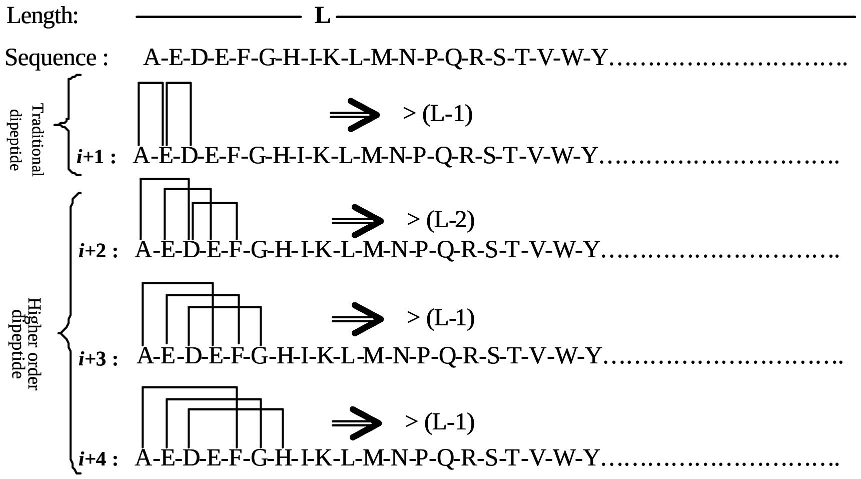
Length:
L
Sequence : A-E-D-E-F-G-H-I-K-L-M-N-P-Q-R-S-T-V-W-Y…………………………
Traditional dipeptide
> (L-1)
i+1 : A-E-D-E-F-G-H-I-K-L-M-N-P-Q-R-S-T-V-W-Y…………………………
> (L-2)
i+2 : A-E-D-E-F-G-H-I-K-L-M-N-P-Q-R-S-T-V-W-Y…………………………
Higher order dipeptide
> (L-1)
i+3 : A-E-D-E-F-G-H-I-K-L-M-N-P-Q-R-S-T-V-W-Y…………………………
> (L-1)
i+4 : A-E-D-E-F-G-H-I-K-L-M-N-P-Q-R-S-T-V-W-Y…………………………

Figure 1

.

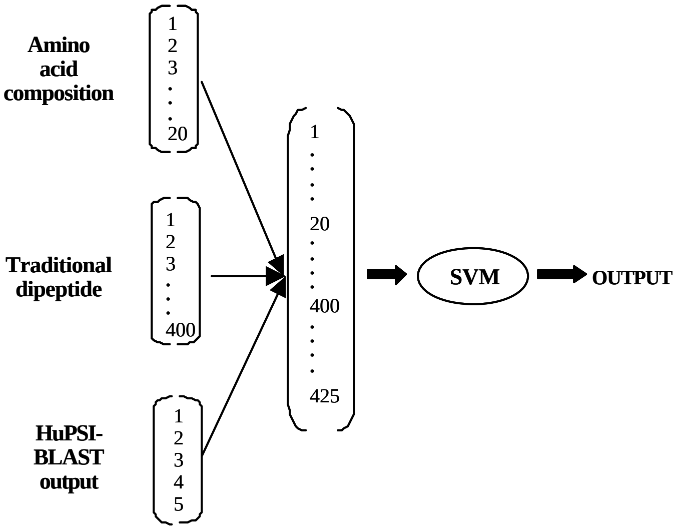
Amino acid composition
1
2
3
20
Traditional dipeptide
1
2
3
400
HuPSI-BLAST output
1
2
3
4
5
1
20
400
425
SVM
OUTPUT

Figure 2a

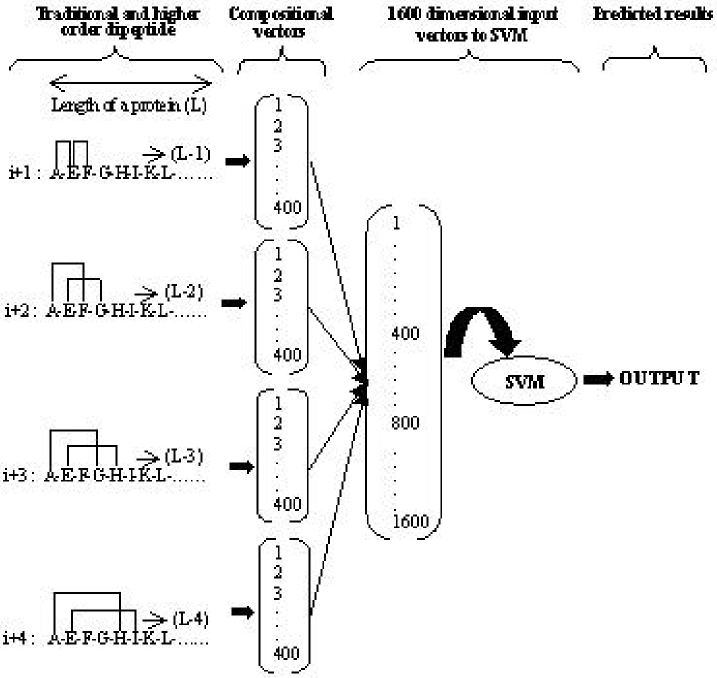
Traditional and higher order dipeptide
Compositional vectors
1600 dimensional input vectors to SVM
Predicted results
Length of a protein (L)
i+1 : A-E-F-G-H-I-K-L-......
(L-1)
i+2 : A-E-F-G-H-I-K-L-......
(L-2)
i+3 : A-E-F-G-H-I-K-L-......
(L-3)
i+4 : A-E-F-G-H-I-K-L-......
(L-4)
1
2
3
400
800
1600
SVM
OUTPUT

Figure 2b

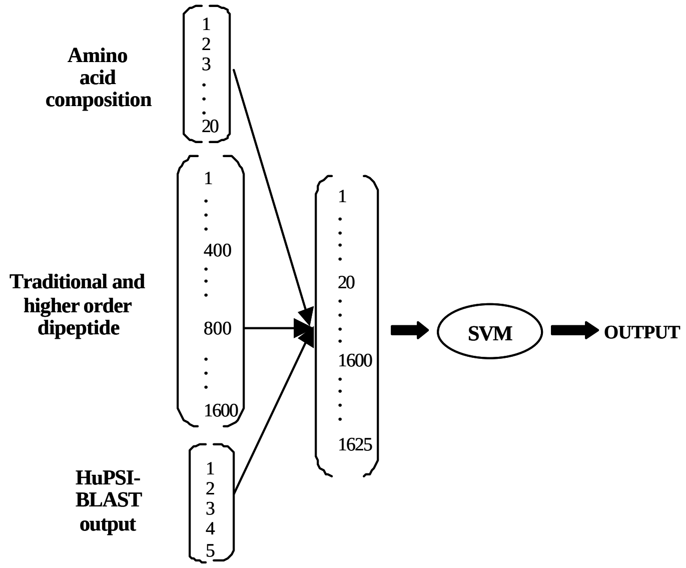
Amino acid composition
1
2
3
20
Traditional and higher order dipeptide
1
400
800
1600
HuPSI-BLAST output
1
2
3
4
5
1
20
1600
1625
SVM
OUTPUT

Figure 2c

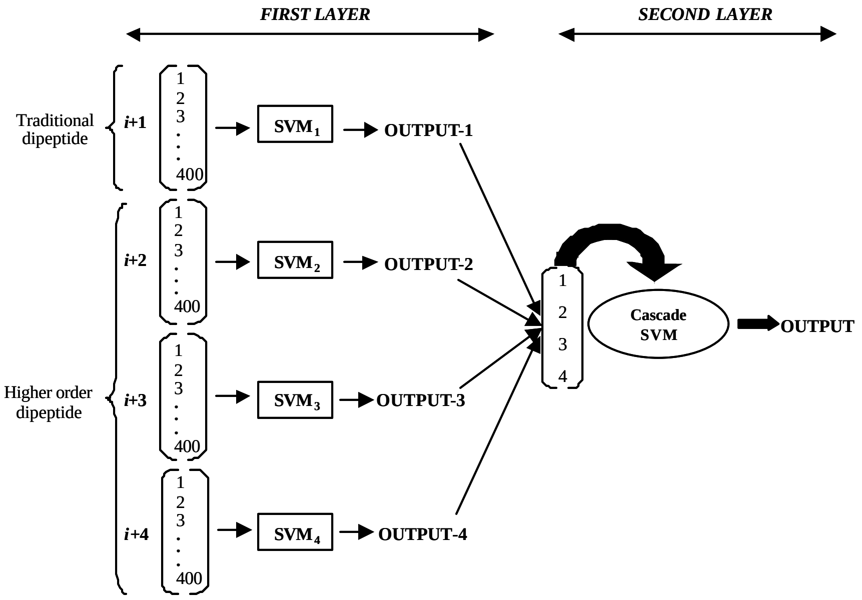
FIRST LAYER
SECOND LAYER
Traditional dipeptide
Higher order dipeptide
i+1
i+2
i+3
i+4
1
2
3
.
.
.
400
SVM1
SVM2
SVM3
SVM4
OUTPUT-1
OUTPUT-2
OUTPUT-3
OUTPUT-4
1
2
3
4
Cascade SVM
OUTPUT

Figure 2d

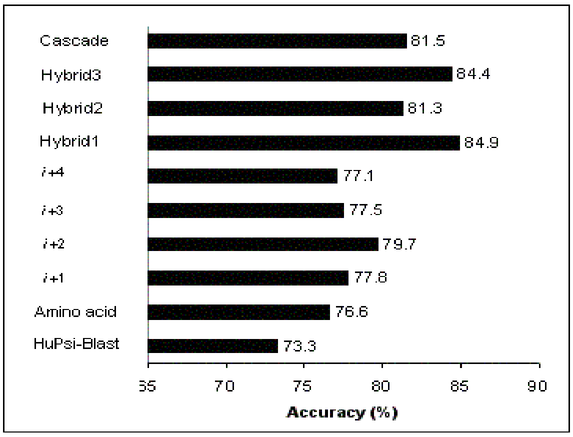
Cascade
Hybrid3
Hybrid2
Hybrid1
i+4
i+3
i+2
i+1
Amino acid
HuPsi-Blast
81.5
84.4
81.3
84.9
77.1
77.5
79.7
77.8
76.6
73.3
65
70
75
80
85
90
Accuracy (%)

Figure 3

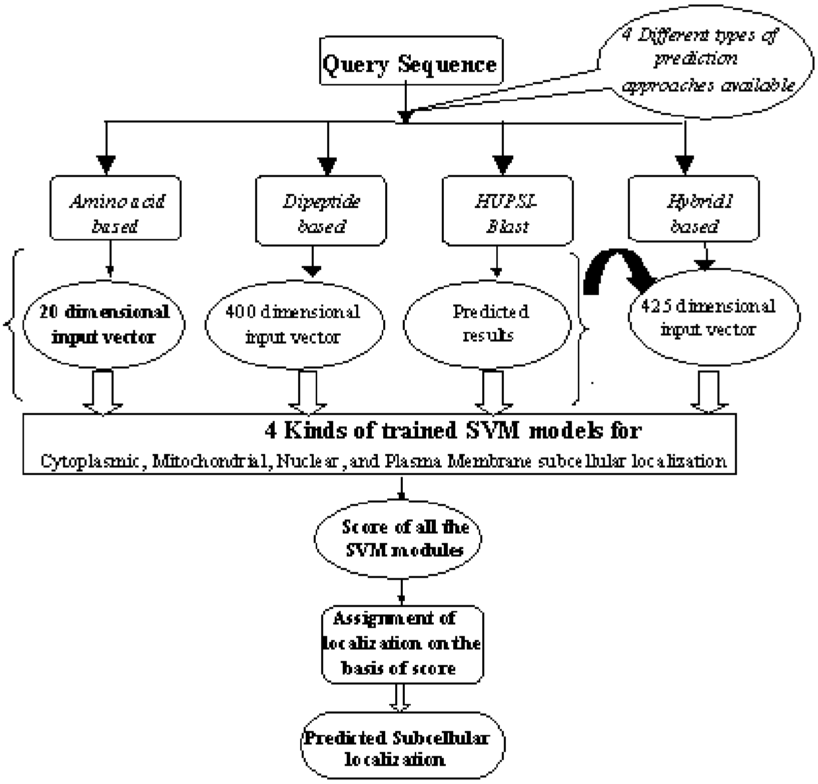
Query Sequence
4 Different types of prediction approaches available
Amino acid based
Dipeptide based
HUPSL-Blast
Hybrid1 based
20 dimensional input vector
400 dimensional input vector
Predicted results
425 dimensional input vector
4 Kinds of trained SVM models for
Cytoplasmic, Mitochondrial, Nuclear, and Plasma Membrane subcellular localization
Score of all the SVM modules
Assignment of localization on the basis of score
Predicted Subcellular localization

Figure 4