A Support Vector Machine Approach for LTP Using Amino Acid Composition

N. Hemalatha and N.K. Narayanan

Abstract Identifying the functional characteristic in new annotated proteins is a challenging problem. With the existing sequence similarity search method like BLAST, scope is limited and accuracy is less. Rather than using sequence information alone, we have explored the usage of several composition, hybrid methods, and machine learning to improve the prediction of lipid-transfer proteins. In this paper, we have discussed an approach for genome wide prediction of LTP proteins in rice genome based on amino acid composition using support vector machine (SVM) algorithm. A predictive accuracy of 100 % was obtained for the module implemented with SVM using polynomial kernel. This approach was compared with an All-plant method comprising of six different plants (wheat, maize, barley, arabidopsis, tomato and soybean) which gave an accuracy of only 70 % for SVM.

Keywords SVM · LTP · All-Plant · Machine learning

1 Introduction

Among the various abiotic stresses that affect the rice production, high temperature is one of the main concerns. Development of high-temperature tolerant rice variety has become a major area for rice scientists to work upon [1].

Lipid-transfer proteins (LTP) are basic 9-kDa proteins. They are present in flowering plants in large amounts and can boost in vitro phospholipids transfer between membranes. They can also bind acyl chains. These properties help them to further participate in membrane biogenesis and also regulation of the intracellular fatty acid pools [2]. Studies by Wang and Liu have shown that the expression of

N. Hemalatha (✉)
AIMIT, St. Aloysius College, Mangalore 575022, India
e-mail: hemasree71@gmail.com

N.K. Narayanan
Department of Information Science Technology, Kannur University, Kannur 670567, India

D.K. Lobiyal et al. (eds.), *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*, Lecture Notes in Electrical Engineering 396, DOI 10.1007/978-81-322-3589-7_2

LTPs can be induced by environmental stresses like extreme temperatures, osmotic pressures, and drought [3, 4].

Rice being one of the major food crops is a subject of research worldwide. Because of the advances in the sequencing techniques in the past few years, whole genomic sequences of rice which includes subspecies *japonica* and *indica* is publicly available. Manual annotation of these sequences is not feasible because data is large. This paper discusses a novel approach of prediction for high-temperature resistant protein LTP using amino acid composition features and hybrid information of proteins. The problem undertaken in this paper is the identification of functional characteristic of newly annotated proteins. Many varieties of *indica*, the subspecies of *Oryza*, are yet to be sequenced and has to be annotated which manually is impossible. Hence, development of prediction approaches will definitely help the biologists for future annotations. The machine-learning approach SVM with three different kernels and nine feature extraction techniques was used. An All-plant model with six different plants (wheat, maize, barley, arabidopsis, tomato, and soybean) using amino acid composition was also created and compared with the new developed method to prove the species-specific property of the classifier.

The paper is organized as follows: Sect. 2 present the data sets to be used in the experiments and steps to extract features from the sequences involved. This section also covers the feature extraction methods applied on the data sets and classification model building. Section 3 presents the performance measures to evaluate SVM with different kernel and feature methods. Comparison of newly developed model in SVM with existing sequence search algorithm PSI-BLAST is discussed in Sect. 4. Section 5 discusses the experimental results over the testing dataset and species-specific property of the developed model. Finally, Sect. 6 concludes with recommendations for future research.

2 Materials and Methods

2.1 Materials

A total of 105 LTPs and non LTP's belonging to both *japonica* and *indica* were collected from Uniprot Knowledgebase (UniProt KB) and National Centre for Biotechnology Information (NCBI). To make the dataset completely nonredundant, CD-HIT software was applied for removing sequences highly similar to other sequences with a threshold of 90 % [5]. Majority of data collected were computationally predicted, and hence to confirm the sequences to be of LTP family Prosite and PFam databases were used. Finally, 105 LTPs were retained for positive dataset and a set of negative samples was constructed from 105 non LTPs from *Oryza sativa*. To check the species-specific property of the approach, another set of negative samples of 105 non LTPs from other plants were taken.

2.2 Methods

Binary SVM Support Vector Machine (SVM) is a classification algorithm based on statistical learning theory. SVM can be applied to pattern classification by mapping the input vectors to a feature space which is of higher dimension [6]. A binary SVM is used in this work to classify sequences into LTP's and non LTP's. Let $s = s_1, s_2 \ldots s_n$ denote a protein sequence of length n where $s_i \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ and dimension $R = R_1, R_2 \ldots R_9$. An ideal mapping for classifying sequences into LTP's and non LTP's from R^9 space into $-1, +1$, where +1 corresponds to LTP class and -1 to non LTP classes, respectively [7].

Let $(r_j, q_j); j = 1, 2 \ldots N$ denote the set of N training sets, where q_j denotes the either class LTP or non LTP, for the input feature vector r_j. Kernel functions are introduced for nonlinearly separable problems as training sets on normalization contain random values and which makes optimization problem more simpler. SVM first maps the input feature vector to a higher dimensional space H with a kernel function k and then is combined linearly with a weight vector w to obtain the output. The binary SVM is trained to classify whether the input protein sequence belongs to the LTP or non LTP class.

SVM develops a discriminant function for classifying LTP by solving the following optimization problem:

$$\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i=1}^{N} y_i y_j k(r_i, r_j) \alpha_i \alpha_j$$

subject to

$$0 \leq \alpha_i, \text{for } i = 1, 2, \ldots n;$$

$$\sum_{i=1}^{N} y_i \alpha_i = 0$$

The kernel function $k(r_i, r_j) = \phi(r_j)^T \phi(r_i)$ and the weight vector $w = \sum_{i=1}^{N} \alpha_i q_i \phi(r_i)$, where ϕ represents the mapping function to a higher dimension and α represents Lagrange multiplier. The optimization gives the values for the parameters α_j and the resulting discriminant function f is given by

$$f(r_i) = \sum_{j=1}^{N} \alpha_j q_j k(r_i, r_j) + b = w^T \phi(r_i) + b$$

where bias b is chosen so that $q_j f(r_j) = 1$ for all j with $0 < \alpha_j < \gamma$. The class corresponding to input pattern r_i is LTP if $f(r_i) \geq 0$ or non LTP if $f(r_i) < 0$.

3 Features

For converting the protein characteristics to feature vectors, effective mathematical expressions should be formulated. This is necessary for applying machine-learning technique that is relevant to the prediction tasks. In this paper, we have used five different composition methods and four hybrid methods which are discussed in the following section.

3.1 Composition-Based Features

Amino acid composition This composition provides a 20-dimensional feature vector. This is an important attribute since this feature denotes a fundamental structural aspect of a protein encapsulating the information regarding the occurrence of each amino acid in the particular protein sequence. The fraction of each amino acid a_i in the given sequence is given by the formula:

$$P(a_i) = \frac{Na_i}{\sum_{i=1}^{20} Na_i} \tag{1}$$

where Na_i represents the total number of a particular amino acid a_i present in the sequence and $\sum_{i=1}^{20} Na_i$ represents the total number of amino acids present in the given protein sequence.

Dipeptide composition This composition provides a 400 (20×20) dimensional vector. This feature encapsulates the local information of each protein sequence utilizing the sequence order effects. The fraction of each dipeptide $a_i a_j$ in the given sequence is given by the formula:

$$P(a_i a_j) = \frac{Na_i a_j}{\sum_{i=1}^{20} \sum_{j=1}^{20} Na_i a_j} \tag{2}$$

where $Na_i a_j$ gives the total number of $a_i a_j$ dipeptides in the sequence and $\sum_{i=1}^{20} \sum_{j=1}^{20} Na_i a_j$ represents the total number of all dipeptides present in the given protein sequence.

Tripeptide composition This composition provides a 8000 (20×400) dimensional vector. This composition gives the properties of the neighboring amino acids The fraction of each tripeptide $a_i a_j a_k$ in the given sequence is given by the formula:

$$P(a_i a_j a_k) = \frac{Na_i a_j a_k}{\sum_{i=1}^{20} \sum_{j=1}^{20} \sum_{k=1}^{20} Na_i a_j a_k} \tag{3}$$

where $Na_i a_j a_k$ gives the total number of $a_i a_j a_k$ and $\sum_{i=1}^{20} \sum_{j=1}^{20} \sum_{k=1}^{20} Na_i a_j a_k$ represents the total number of all tripeptides present in the given protein sequence.

3.2 Four-Parts Composition

This composition is based on the assumption that different parts of a sequence can provide valuable information. In this composition, query sequence is divided into four fragments of equal length and amino acid composition from each fragments are calculated separately using Eq. (1). Each fragment of 20 dimensions gets concatenated to form 80 dimensional feature vector.

3.3 Three Parts Composition

This composition is also otherwise called as terminal-based N-center-C composition. This determines the signal peptides at N or C terminal region of different proteins. To identify these signal peptides, amino acid composition has to be calculated separately from the terminals N and C and remaining from the center region. For each region a 20-dimensional vector will be created using Eq. (1), so for the three regions the combined feature will have a dimension vector of 60.

3.4 Hybrid-Based Features

Hybrid1 approach This approach combines amino acid and dipeptide composition features of a protein sequence and is calculated using Eqs. (1) and (2), respectively. Because amino acid and dipeptide compositions are combined, feature vector will have a dimension of 420, i.e., 20 for amino acid and 400 for dipeptide.

Hybrid2 approach This approach combines amino acid and tripeptide composition features of a protein sequence using Eqs. (1) and (3), respectively. This approach has a feature dimension of 8020, i.e., 20 for amino acid and 8000 for tripeptide.

Hybrid3 approach In this approach, amino acid was combined with four part composition which was calculated using Eq. (1) and has a dimension of 100.

Hybrid4 approach In this approach, we combined amino acid composition calculated using Eq. (1), dipeptide calculated using Eq. (2) and three parts calculated using Eq. (1) to have an input feature dimension of 480 (20 for amino acid, 400 for dipeptide and 60 for three parts composition).

4 Performance Measure

Some standard evaluation methods are used to measure the performance of the algorithm. The two methods which have been used for this purpose are cross validation and independent data test. Cross-validation techniques can be used to test the predictive performance of models as well as to help prevent a model being over fitted. This technique can be of various folds like 10-fold, 20-fold, etc. In the *k*th fold cross validation, the data set is divided into k subsets and each subsets contains equal number of proteins. The k subsets are then grouped into $(k-1)$ training set and remaining one as testing set. This procedure is repeated k times so that every subset is at least used once for testing. In the independent dataset test, testing dataset is totally independent of the training set. Selection of data for training and testing are independent of each other.

The standard evaluation metric used are sensitivity (Sn), specificity (Sp), precision (Pr), accuracy (Acc), F-measure (F), and Mathew correlation coefficient (MCC). Actual prediction of positive and negative data of LTP is measured by sensitivity and specificity respectively. Precision defines the proportion of the predicted positive cases of correct LTP. Accuracy measures the proportion of the total number of correct predictions of LTP. Recall calculates the correctly identified proportion of positive cases of LTP. MCC is used especially when number of positive and negative data differs too much from each other. The value of MCC ranges between −1 and 1 and a positive value indicates a better prediction performance. TP, FP, TN, FN are the numbers of true positives, false positives, true negatives, and false negatives, respectively. Following are the equations used for their calculations:

$$\text{Sensitivity} = \left(\frac{\text{TP}}{\text{TP}+\text{FN}}\right) \times 100 \tag{4}$$

$$\text{Specificity} = \left(\frac{\text{TN}}{\text{FP}+\text{TN}}\right) \times 100 \tag{5}$$

$$\text{F-measure} = \left(\frac{2 \times \text{Pr} \times \text{Sn}}{\text{Pr}+\text{Sn}}\right) \times 100 \tag{6}$$

$$\text{Accuracy} = \left(\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}\right) \times 100 \tag{7}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{8}$$

5 Similarity Search

In sequence similarity searching, a query sequence is searched against sequence databases using alignment which can be statistically assessed. This information can be used to infer homology and transfer information to the query sequence about its match with the database. This similarity between sequences can give clues to similarities in molecular structure and function as well to discover evolutionary relationships between the sequences.

PSI-BLAST tool is widely used in bioinformatics for sequence similarity search. This tool was used for sequence similarity search to find the similarity of the given sequences with other related sequences. This tool compared a protein sequence with a created database [8]. In this paper this tool is used for a comparative study of search using PSI-BLAST to the one presented in this paper using the classifier discussed in Sect. 2.

6 Experiments and Results

Binary SVM was implemented using SVM^{light} [9] which is known to be a fast optimization algorithm. We use a tenfold cross validation and independent data test with different types of kernels to evaluate the accuracy in the LTP classification. The kernel type and parameters were set based on best accuracy.

6.1 Statistical Tests of SVM Classifiers

Applying independent data test in SVM with three different kernels for a total of nine different compositions, an accuracy of 100 % was obtained for amino acid, dipep, and tripep compositions with linear kernel (Table 1). Applying 10-fold cross-validation test in SVM, 98 % accuracy was obtained for four-parts and hybrid1 composition with linear kernel (Table 1).

Amino acid involves less number of features and less complexity, hence we considered amino acid as the best composition. Also linear kernel is the most simplest among the three kernels in SVM. Hence for this classifier, we have chosen amino acid composition with linear kernel which has only 20 dimension obtained for independent test. Cross-validation result was not considered because it could obtain only 98 % for four-parts and hybrid1 composition whose feature size is 80 and 420 which is much higher than amino acid composition. The performance comparison of SVM is depicted in the Fig. 1.

Table 1 Validation of Independent and cross-validation test results with SVM

Composition	Independent test			Cross-validation test		
	Acc	F	MCC	Acc	F	MCC
Aminoacid	100	100	1	90	94	0.85
Dipeptide	100	100	1	96	98	0.92
Tripep	100	100	1	53	53	0
3-parts	80	88	0.61	89	11	0.73
4-parts	95	100	0.9	98	97	0.95
Hybrid1	50	50	0	98	97	0.95
Hybrid2	50	50	0	96	0	0.91
Hybrid3	95	100	0.9	92	92	0.81
Hybrid4	85	100	0.73	52	19	0.11

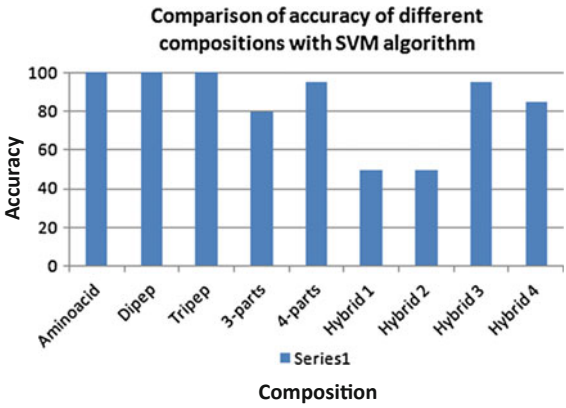

Fig. 1 Performance comparison of accuracy of 9 different compositions

6.2 *Similarity Search*

PSI-BLAST, the sequence similarity search tool, was compared to the newly developed approach. For this, 10-fold cross validation was conducted for similarity search tool which generated a very less accuracy of 67.6 % Table 2. This result shows that similarity search is not an efficient tool for comparison compared to feature based approach.

6.3 *Species-Specific Classifier*

To find what happens to the classifier when non-rice LTP patterns are included in the training set, two tests were conducted. In the first, an All-plant method consisting of six plants namely arabidopsis, wheat, maize, barley, tomato, and soybean

Table 2 Prediction result of LTP with similarity search (tenfold cross validation used)

Test	No. of Test sets	No. of correctly predicted	Average
1	19	12	63
2	19	13	68
3	19	14	73.6
4	19	12	63
5	19	13	68
6	19	14	73.6
7	19	13	68
8	19	14	73.6
9	19	12	63
10	8	5	62.5
			67.6

was developed and compared to the rice-specific classifier. In the second test, the newly developed method was tested with the above six plants.

Comparison with All-plant method In this test, an All-plant method was developed in SVM using all the three kernels. For creating this, six plants were taken namely arabidopsis (*Arabidopsis thaliana*), wheat (*Triticum aestivum*), maize (*Zea mays*), barley (*Hordeum vulgare*), tomato (*Solanum lycopersicum*) and soybean (*Glycine max*), including a total of 174 data in the training set. In the case of newly developed All-plant model, we have used the simple amino acid approach, which was having an accuracy of 100 % for rice-specific classifier.

On comparison of newly created rice-specific classifier with corresponding All-plant module based on the rice independent training set, the former showed an increase of 50 % accuracy with respect to linear kernel. From Table 3, it can be seen that All-plant method is having 70 % accuracy for both polynomial and RBF kernel and only 50 % for linear kernel. These results clearly indicate the advantage of species-specific classifier. Methodology for creating both the model were identical, i.e., have used the amino acid composition approach. This strongly suggests that species-specific prediction systems are much better compared to general ones.

Performance on other Plants In the second validation, we checked the performance of newly developed method on six plants namely arabidopsis, wheat, maize, barley, soybean, and tomato. The results obtained are tabulated in Table 4. The result of SVM model with RBF kernel revealed accuracy of 96 % for four

Table 3 Comparison of All-plant model with new model

Method	Algorithm	Sn	Sp	Acc	MCC
All-plant	Linear	100	0	50	0.00
	Polynomial	40	100	70	0.50
	RBF	40	100	70	0.50
New model SVM	Linear	100	100	100	1.00
	Polynomial	100	100	100	1.00
	RBF	100	100	100	1.00

Table 4 Performance of SVM model on six plants

Plants	Linear	Poly	RBF
	Acc	Acc	Acc
Arabidopsis	60	64	96
Maize	50	92	92
Wheat	60	96	96
Barley	60	64	95
Tomato	60	88	96
Soyabean	60	96	95.8

plants namely arabidopsis, tomato, soybean, and wheat whereas for other two remaining plants, i.e., barley and maize, the prediction accuracy was equal or below 95 %. However, when the same model was run only on rice proteins, the new model achieved 100 % accuracy during independent data test. This difference between the performances of both the method with rice and other plants indicate that there might be some species-specific feature in rice dependent method.

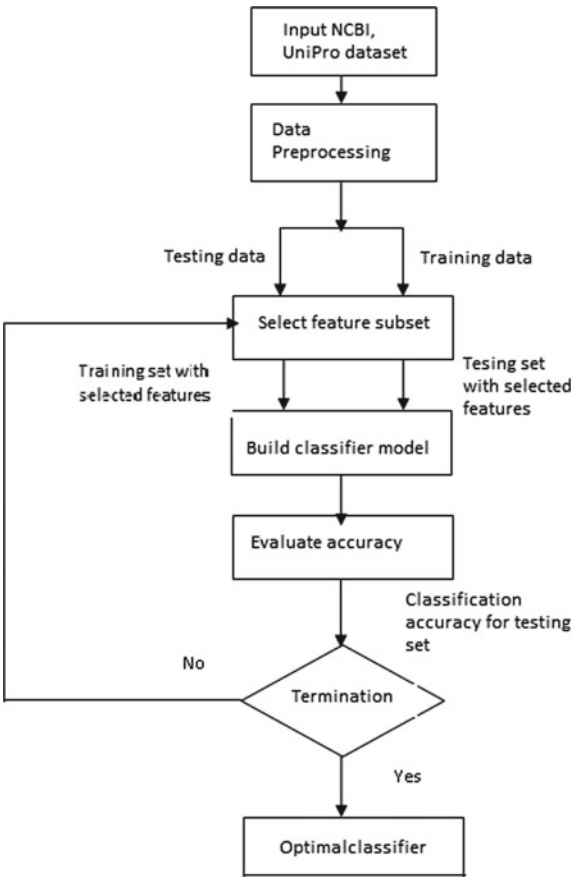

Fig. 2 Architecture of new model using SVM

6.4 Description of Architecture

The overall architecture of the methodology used for developing new method using SVM is depicted in the Fig. 2.

7 Conclusions

The use of computational tools and web databases has promoted the identification of various functional proteins. The different computational methods currently available are general ones and can be used for functional annotation of a given protein by determining their prediction accuracy. The sequencing of varieties of *oryza sativa indica* subspecies of *oryza sativa*, commonly used in Asian countries is yet to be completed and these newly sequenced proteins will have to be annotated. Different stress prediction tool for rice according to our knowledge is unavailable and the general methods are available but less accurate. Here, we have proposed a highly accurate prediction algorithm using SVM for identification of LTPs in *Oryza sativa*. Also in this work, we have proved the advantage of species-specific classifier over the general ones. This further substantiates that the new method using SVM will contribute significantly to the various annotation projects in rice. In our future work on LTPs, we plan to further elaborate the database of LTPs in rice and also study the effect of various other proteins with respect to this abiotic stress.

References

1. Krishnan, P., Ramakrishnan, B., Reddy, K. R., Reddy, V. R.: Chapter three-High-Temperature Effects on Rice Growth, Yield, and Grain Quality. Adv. Agron. 111, 87–206 (2011).
2. Kader, J. C.: Lipid-transfer proteins in plants. Annu. Rev. Plant. Biol. 47(1),627–654 (1996).
3. Wang, N. J., Lee, C. C., Cheng, C. S., Lo, W. C., Yang, Y. F., Chen, M. N., Lyu, P. C.: Construction and analysis of a plant non-specific lipid transfer protein database (nsLTPDB). BMC genomics. 13(Suppl 1) (2012).
4. Liu, Qiang, Yong Zhang, Shouyi Chen.: Plant protein kinase genes induced by drought, high salt and cold stresses. Chinese Sci Bull. 45(13), 1153–1157 (2000).
5. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13,1658–1659 (2006).
6. Vapnik, V.: The nature of statistical learning theory. Springer Science & Business Media. (2000).
7. Ma, J., Nguyen, M. N., Rajapakse, J. C.: Gene classification using codon usage and support vector machines. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 6(1),134–143 (2009).
8. Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids res. 25(17), 3389–3402 (1997).
9. Schlkopf, B., Burges, C. J. Advances in kernel methods: support vector learning. MIT press (1999).

http://www.springer.com/978-81-322-3587-3

Proceedings of the International Conference on Signal, Networks, Computing, and Systems
ICSNCS 2016, Volume 2
Lobiyal, D.K.; Mohapatra, D.P.; Nagar, A.K.; Sahoo, M.N. (Eds.)
2016, XVI, 348 p. 156 illus., Hardcover
ISBN: 978-81-322-3587-3