Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2, No.5 : 2024
ISSN : **1906-9685**

SEQUENCE-BASED HEAT SHOCK PROTEIN PREDICTION IN PLANTS USING MACHINE LEARNING

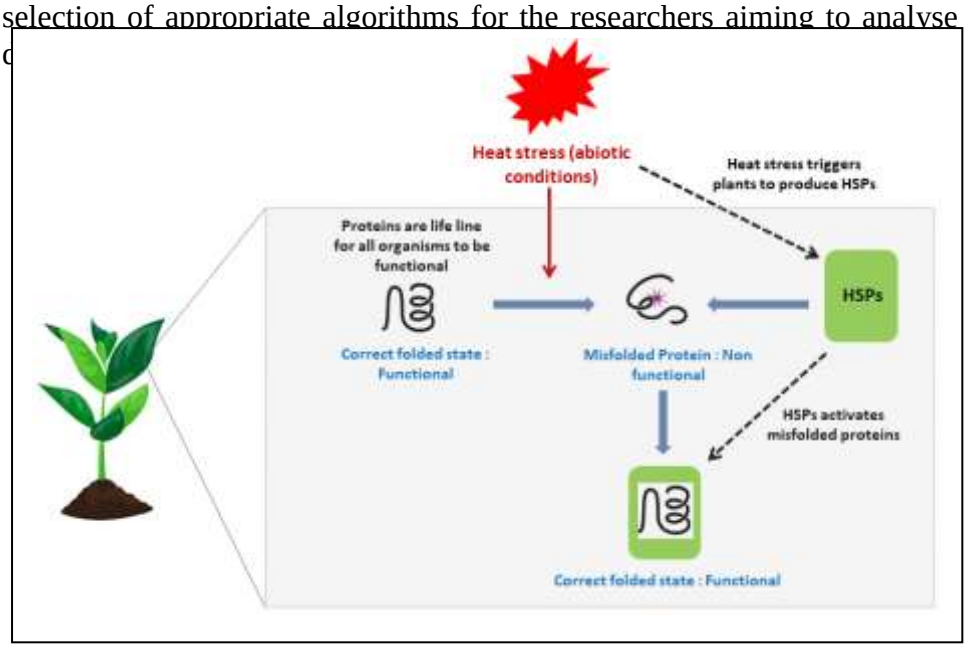
A.Amuthavalli , Assistant Professor Department of Artificial Intelligence and Machine Learning Sri Krishna Adithya College of Arts and Science
Dr.K.Nandhini, Assistant Professor Department of Computer Applications (PG) Vels Institute of Science, Technology & Advanced Studies (VISTAS)

Abstract
Heat shock proteins (HSPs) plays as a guardian in plants by defending and maintaining growth in response to various abiotic stress conditions. Plants are sessile in nature and thus need to constantly adjust to various environmental conditions; any interaction with abiotic or biotic stress conditions will lead to its reduced functioning like RNA, DNA or protein synthesis. HSPs are group of proteins and act as molecular guardians in plants under stress by enabling and keeping proteins in correct folding and maintain cellular functionality. With advent of sequencing technology huge volume of plant sequence data has been published without much of annotation. Although HSPs play key role in plants, their computational analysis remains limited. There is need for developing efficient Machine Learning (ML) tools for analysing and identification of crucial HSPs genes and its characterization. In this study the proposed approach used for prediction of HSP versus non HSP. The feature selection procedures Amino Acid Composition (AAC), Dipeptide Composition (DC) and Composition Transition and Distribution (CTD) applied for 6445 protein sequences with 567 features were used for classification. The random forest algorithm produced the best results of 91% out of all the algorithms examined. Cross-validation and independent data set validations were applied to test the predictability and performance of the proposed algorithm. The results revealed that the proposed approach might be highly useful in predicting HSP computationally.
Keywords
Machine learning, Classification, Random forest, Naïve bayes, KNN.

Introduction

Heat shock proteins (HSPs) are omnipresent in nature and acts as molecular on maintaining function of proteins by keeping it in correct folding structure [1][2]. HSPs participate in many molecular developments like assembly of proteins, secretion and transportation by constructing stable protein structure functioning properly in a cell [3][4]. Elevated temperature condition is one of abiotic stress, which leads to decreased plant performance and in turn crop yield word wide [5]. HSP helps in preventing plants suffering from Heat stress has depicted in figure 1. Plants attempt to effectively cope up with the high temperature stress by actively producing HSPs. During stress conditions, proteins normally synthesized chaperones in the cells will undergo modifications like aggregations, denaturation which leads to inactivation and in turn cell damage[6][7]. However HSPs will protect the cells by either removing inactive proteins or making it soluble form and reactivates its functionality [8]. Production of different types of HSPs expressed in a cell varies between plant species and the environmental conditions [9]. Investigation of different types of HSPs and its role in combating drought situations is active area of research across the world laboratories [10][11]. In order to minimize the cost and laborious experiments on characterizing the HSPs from deluge of published data it will be prudent to first predict and analyse Heat shock proteins through various computational approaches. Conventional computational methodologies like similarity based search

tools NCBI or EXPASY BLAST were tested for identifying Heat shock proteins[12][13]. However these strategies did not lead to precise characterization of HSP proteins and its types as the query data did not have significant sequence match .Other non-sequential based machine learning approaches were carried out by different research groups primarily based on support vector machine (SVM) using amino acid and di peptide compositions[14][15][16]. The amino acid composition in split mode and dipeptide feature mode applied to predict HSPs with overall 90.7% accuracy rate [17]. Heat shock protein predictor called "iHSP-RAAAC" was developed using amino acid alphabet as a feature tool and with jackknife method obtained 87.42% predictive accuracy [18]. Convolutional neural network based tool was developed to validate different type of Heat shock proteins and demonstrated that measure of test accuracy (F1 scores) increased by 10-20% [19]. However there are only limited published data available for analysing plant specific HSPs. Recently, Naïve Bayes algorithm was applied to classify different types of HSPs from plants based on AAC and DC [20]. In this study, a method is developed by applying three different types of machine learning approaches to assess the performance of HSPs prediction using established models. Our study would provide selection of appropriate algorithms for the researchers aiming to analyse specific HSPs from plant c



II.MATERIALS AND METHODS

A. Datasets

The protein sequences have been downloaded from https://www.uniprot.org/ for various plants. For the construction of positive and negative datasets 3908 HSPs were utilised as positive data, and 2537 HSPs were used as negative data.

B. Methodology

The non-redundant positive dataset and negative dataset of various plants protein sequences were collected from the uniprot database, the features has extracted from the various feature selection technique like a AAC, DC and CTD. The extracted protein features were computed and it was applied as input data for classification model building. The machine learning models random forest, naïve bayes and KNN were used to select the best model for HSP versus non HSP classification. The classification models were compared based on the performance measures. Figure 2 illustrates the prediction model procedure.

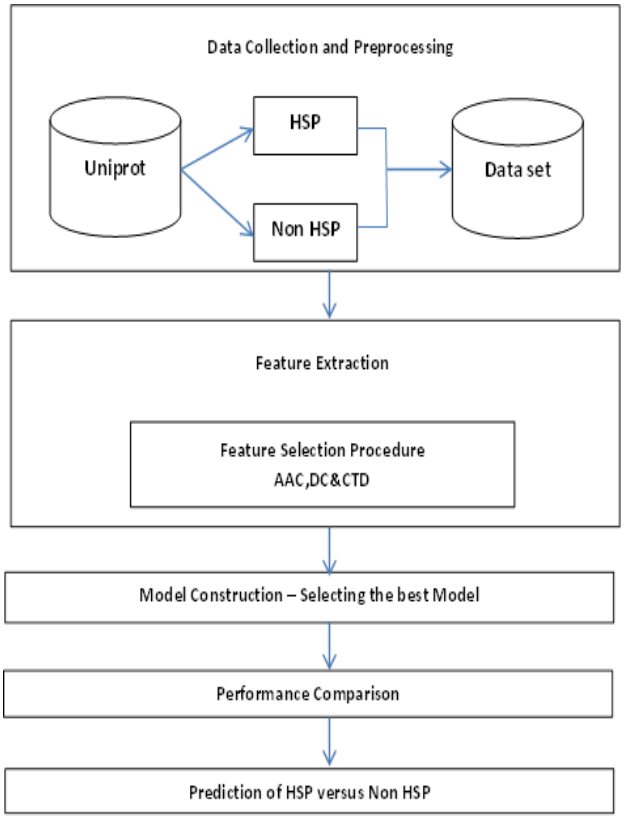


Figure 2: Machine Learning Model

C. Feature

The accurate prediction of HSPs depends on the appropriate selection of classifier and relevant set of parameters applied for evaluation. Three different extraction methods specific for proteins such as AAC, DC and CTD were computed and used as input data for the construction of classification models. The python package PyBioMed is used for feature extraction of amino acids and peptide from a protein dataset. Totally 567 features obtained through features of AAC (20), DC (20*20) and CTD (147) were tested based on amino acids, dipeptide and physiochemical properties.

Amino Acid Composition (AAC)

A series of 20 amino acids constitutes a given protein composition. The composition technique is used for determining the characteristics of individual amino acid in a given protein sequence. Amino acids are represented by a 20-dimensional vector in this manner. The following equations were used to assess the property of amino acid character in the protein sequence. A protein sequence represented by 'Pro' and length number 'Num' can be characterized as a sequence y1, y2, y3….y_n, where y_1, y_2,…. y_n are the amino acids. The result comprises the existence of every amino acid in a sequence

$$\text{AAC of Xi} = \text{Quantity of occurrences of Xi in Pro} / \text{Num}$$

Example

Protein_sequence="ADGCGVGEGTGQGPMCNCMCMKWVYADEDAADLESDSFADEDASLESDSFPWSNQRVFCSFADEDAS"

The occurrences of individual amino acid in a given protein sequences are

{'A': 11.94, 'C': 7.463, 'E': 8.955, 'D': 14.925, 'G': 8.955, 'F': 5.97, 'I': 0.0, 'H': 0.0, 'K': 1.493, 'M': 4.478, 'L': 2.985, 'N': 2.985, 'Q': 2.985, 'P': 2.985, 'S': 11.94, 'R': 1.493, 'T': 1.493, 'W': 2.985, 'V': 4.478, 'Y': 1.493}

Dipeptide Composition (DC)

The dipeptide approach was used to capture the comprehensive information and specific patterns of each protein using sequence order effects. This methodology utilizes 400 (20 x 20) dimensional vectors of a protein sequence. The equations were used to evaluate the property and nature of an amino acid in a given protein sequence.

$$\text{AAC of XiXj} = \text{Number of occurrences of XiXj in Pro} / \text{Num}$$

For all 1< i,j<=20 .

Composition Transition and Distribution (CTD)

CTD constitutes amino acid properties such as hydrophilicity, mass, hydrophobicity, polarity, charge, solvent solubility, secondary and tertiary structure. All these parameters were utilised to identify characteristics for the classification model .For categorization, 147 descriptors were developed for a specific protein sequence.

D. Machine learning algorithms

ML methodologies like Naïve Bayes, Random Forest (RF), and *k*-nearest neighbors (KNN) are supervised algorithms commonly applied for the classification of test data or for predicting numerical based regressed trait values which needs defined labels [21][22].Even though these three MLs are unique in nature, they have common characteristics in which each algorithm analyse the data over a feature space with transformed version and attempt to give best answers to the problems with minimized empirical risk [23]

Random Forest

The Random Forest is commonly applied to specifically solve problems associated with regression and classification of data. It is an ensemble classifier embedded with different decision trees and primarily works on conquers and divide strategy for better performance output on identifying variables of interest from huge datasets [24][25]

Naïve Bayes

Nave Bayes is well established probabilistic classifier based on Bayesian theory in machine learning methods. It has robust 'naïve' hypothesis for the independence assumptions between selected features and class variables leads to precise prediction of data samples [26][27]

KNN

KNN is often used for classification of binary or multi class test data or identification of numerical trait values (regression) and needs clear information on labels. KNN is a non-assumption model on underlying data and thus mainly applied for data with irregular decision boundaries or with many prototype classes [28][29]

III RESULTS AND DISCUSSION

A. Performance evaluation

The data set contains 6445 sequences. The training and testing and samples includes 5156 and 1289 respectively. The following processes were implemented to test the efficiency of the model by adopting metrics like Specificity (Spe) , Sensitivity (Sen), precision (Pre), F-measure (F_M) and Accuracy(Acc) are expressed in the following formula.

$$\text{Sen}=\left(\frac{TP}{TP+FN}\right) * 100$$

$$\text{Spe} =\left(\frac{TN}{FP+TN}\right) * 100$$

$$\text{F\_M} =\left(\frac{2*Pr*Sn}{Pr+Sn}\right) * 100$$

$$\text{Pre}= \left(\frac{TP}{FP+TP}\right) * 100$$

$$\text{Acc}=\left(\frac{TP+TN}{TP+FN+TN+FP}\right) * 100$$

B. Ten-fold Cross-Validation (10 fold CV)

Cross-validation method is the most prevalent approaches for evaluating a model. In this model the data is separated into ten subsets, the holdout method is repeated ten times in the ten-fold classification method. Nine subsets are combined to create training data, and a subset is evaluated as a test set each time. It is determined the average of the total number of errors across all trials. In this

strategy, each test uses 90 percent of the total data for training. Among all the classification algorithm the naïve bayes predicts with high accuracy of 97%.

C. Receiver Operating Characteristic (ROC) Curve

The performance of classifier and the nature of the amino acid and dipeptide patterns in a protein data set can be best tested by receiver operating characteristic curve (ROC curve). Basically the output of ROC is projected as a graph that shows how well a classification model performs across all categorization levels. In this curve, two parameters are designed: Degrees of True and False Positives. In this study, the ROC of random forest algorithm has the 98% in combined features selection method.

Predicting with individual features

An self-determining set of data was implemented to measure the overall performance of all methods. Table 1 shows the individual performance of each of the three feature extraction methods. Out of all methods the random forest algorithm has 89 percent accuracy was obtained for AAC, 90 percent for DC and CTD. The random forest algorithm predicts the HSPs with more accuracy. Naïve Bayes algorithm predicts HSP versus non HSP sequences with 74% accuracy in DC.

Method	Algorithm	Precision	Sensitivity	F-measure	Accuracy	10 fold CV	ROC
AAC	Random Forest	0.896	0.936	0.915	0.894	0.891	0.95
	Naïve Bayes	0.75	0.77	0.76	0.701	0.7	0.96
	KNN	0.783	0.895	0.836	0.784	0.78	0.95
DC	Random Forest	0.898	0.961	0.929	0.909	0.913	0.971
	Naïve Bayes	0.835	0.731	0.78	0.746	0.76	0.83
	KNN	0.824	0.918	0.869	0.829	0.80	0.895
CTD	Random Forest	0.893	0.951	0.921	0.9	0.883	0.972
	Naïve Bayes	0.676	0.058	0.107	0.404	0.412	0.971
	KNN	0.704	0.951	0.809	0.724	0.70	0.77

Table 1 Prediction of HSP with individual features

Predicting with combined features

The performance of the proposed system with regard to predictions done on the three algorithms is shown in Table 2 and plotted in figure3. The accuracy of the predictions was 91%, 70% and 72% for random forest, naïve Bayes and KNN. In the 10 fold CV naïve Bayes classifiers has the highest prediction.

Algorithm	Precision	Sensitivity	F-measure	Accuracy	10 fold CV	ROC
Random Forest	0.911	0.96	0.935	0.918	0.91	0.98
Naïve Bayes	0.893	0.582	0.705	0.701	0.97	0.73
KNN	0.705	0.951	0.81	0.725	0.71	0.97

Table 2 Prediction of HSP with combination of features

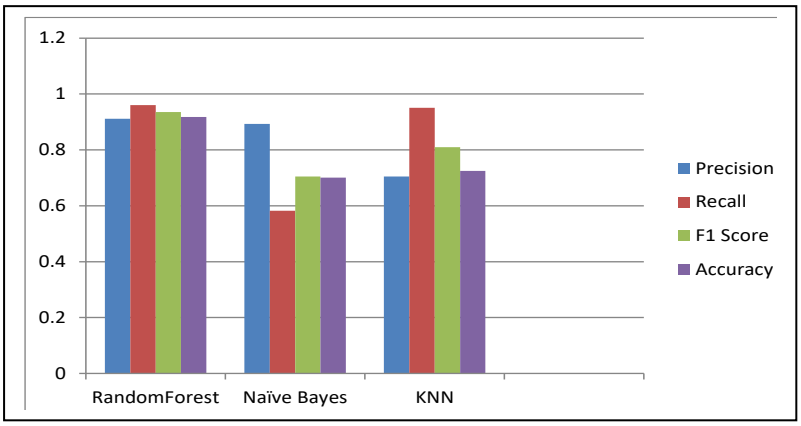


Figure 3: Performance metrics with combination of features

IV CONCLUSION

In this study, specific protein identification such as plant HSPs from large volume of data using supervised algorithm models based on random forest, naïve bayes and KNN are projected .The proposed system describes a machine learning approach for predicting plant HSPs with testing features like amino acid, physiochemical properties and dipeptide composition. In order to overcome imbalanced output and false positives, significantly large dataset of 6445 protein sequences with 567 attributes were used to predict the HSP versus non HSP with the combined features of AAC, DC and CTD. Out of three evaluated machine learning approaches, the random forest algorithm gives the accuracy of 91% in 10 fold CV and 98% in ROC. The outcome of this study is expected to augment the current and future studies in predicting HSPs from plant species.

References

[1] Zeng, X. C., Bhasin, S., Wu, X., Lee, J. G., Maffi, S., Nichols, C. J., et al.Hsp70 dynamics in vivo: effect of heat shock and protein aggregation. J. CellSci. 2004; 117, 4991–5000.

[2] Poulain, P., Gelly, J. C., and Flatters, D. Detection and architecture of small heat shock protein monomers.2010; PLoS ONE 5:e9990.

[3] Tytell, M., and Hooper, P. L. Heat shock proteins: new keys to the development of cytoprotective therapies. Expert Opin. Ther. Targets. 2001; 5, 267–287.

[4] Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, Bruford EA, et al. Guidelines for the nomenclature of the human heat shock proteins. Cell Stress and Chaperones. 2009; 14(1):105–111.

[5] Al-Whaibi, M. H. Plant heat-shock proteins: a mini review. 2011; J. King Saud Univ. Sci. 23, 139–150.

[6] Parsell DA, Lindquist S. The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. Annual Review of Genetics. 1993;27(1):437-496.

[7] Saini J, and Sharma, P. K. Clinical, prognostic and therapeutic significance of heat shock proteins in cancer. Curr. Drug Targets. 2017;19(13):1478-1490.

[8] Kampinga, H. H., and Bergink, S. Heat shock proteins as potential targets for protective strategies in neurodegeneration. Lancet Neurol. 2016; 15, 748–759.

[9] Wu, J., Liu, T., Rios, Z., Mei, Q., Lin, X., and Cao, S. Heat shock proteins and cancer. Trends Pharmacol. Sci. 2017; 38, 226–256.

[10] Lange, O. F., Rossi, P., Sgourakis, N. G., Song, Y., Lee, H.W., Aramini, J. M., et al. Determination of solution structures of proteins up to 40 kDa using CSRosetta with sparse NMR data from deuterated samples. Proc. Natl. Acad. Sci.U.S.A. 2012; 109, 10873–10878.

[11] Kumar, R., Kumari, B., and Kumar, M. PredHSP: sequence based proteome-wide heat shock protein prediction and classification tool to unlock the stress biology. PLoS ONE. 206; 11:e0155872.
[12] Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Quarterly Reviews of Biophysics. 2003; 36(03):307-340.
[13] Ratheesh K, N SN, S PA, Sinha D, Veedin Rajan VB, Esthaki VK, et al. HSPIR: a manually annotated heat shock protein information resource. Bioinformatics. 2012; 28(21):2853–2855.
[14]Volpato V, Adelfio A, Pollastri G. Accurate prediction of protein enzymatic class by N-to-1 Neural Networks. BMC Bioinformatics. 2013; 14(1):1.
[15] Waters ER, Aevermann BD, Sanders-Reed Z.Comparative analysis of the small heat shock proteins in three angiosperm genomes identifies new subfamilies and reveals diverse evolutionary patterns. Cell Stress and Chaperones. 2008; 13(2):127-142.
[16] Zhao XW, Li XT, Ma ZQ, Yin MH. Identify DNAbinding proteins with optimal Chou's amino acid composition. Protein and Peptide Letters. 2012;19(4):398-405.
[17] Ahmad S, Kabir M, Hayat M. Identification of Heat Shock Protein families and J-protein types by incorporating Dipeptide Composition into Chou's general PseAAC. Computer methods and programs in biomedicine.2015; 122(2):165–174.
[18] Feng PM, Chen W, Lin H, Chou KC. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analytical Biochemistry. 2013; 442(1):118–125.
[19] Prabina K. Meher, Tanmaya K. Sahu, Shachi Gahoi, Atmakuri R. Rao. ir-HSP: Improved Recognition of Heat Shock Proteins, Their Families and Sub-types Based On g-Spaced Di-peptide Features and Support Vector Machine. Front Genet.2017;8: 235.
[20] Radhika. Prediction of heat shock proteins in plants based on amino acid composition and machine learning methods Journal of Pharmacognosy and Phytochemistry 2019; 8(3): 3537-3544.
[21] Cai YD, Liu XJ, Xu XB, Zhou GP. Support vector machines for predicting protein structural class. BMC Bioinformatics. 2001; 2(1):1.
[22] Chaurasiya M, Chandulal GB, Misra K, Chaurasiya VK. Nearest-neighbor classifier as a tool for classification of protein families. Bioinformation. 2010; 4(9):396-398.
[23] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–1659.
[24] Breiman, L. Random forests. Mach. Learn. 2001; 45, 5–32.
[25] Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., and Elhaj, F. Feature selection using information gain for improved structural-based alert correlation. PLoS ONE.2016; 11:e0166017.
[26] Frank E, Hall M, Pfahringer B: Locally weighted naive bayes. Morgan Kaufmann Publishers Inc. 2002; 249-256.
[27] Habier, D, R L Fernando, K Kizilkaya, and D J Garrick. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011; 12: 186.
[28] Fawcett, T. An introduction to ROC analysis. Pattern Recog. Lett. 2006; 27,861–874.
[29] Mahood EH, Kruse LH, Moghe GD. Machine learning: A powerful tool for gene function prediction in plants. Appl Plant Sci. 2020 Jul 28;8(7):e11376.
[30] Hu, X., Van Marion, D. M. S., Wiersma, M., Zhang, D., and Brundel, B. J.J. M. The protective role of small heat shock proteins in cardiac diseases: key role in atrial fibrillation. Cell Stress Chaperones. 2017; 22, 665–674.
[31] Nasedkin, A., Marcellini, M., Religa, T. L., Freund, S. M., Menzel,A., Fersht, A. R., et al. Deconvoluting protein (un) folding structural ensembles using X-ray scattering, nuclear magnetic resonance spectroscopy and molecular dynamics simulation. PLoS ONE. 2015; 10:e0125662.
[32] Nagarajan NS, Arunraj SP, Sinha D, Rajan VBV, Esthaki VK, D'Silva P. HSPIR: a manually annotated heat shock protein information resource. Bioinformatics. 2012; 28(21):2853-2855.
[33] Chaurasiya M, Chandulal GB, Misra K, Chaurasiya VK. Nearest-neighbor classifier as a tool for classification of protein families. Bioinformation. 2010; 4(9):396-398.
[34] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001; 43(3):246-255.

[35] Strobl C, Boulesteix AL, Zeileis A, Hothorn T: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007, 8:25