Journal of Pharmacognosy and Phytochemistry

Available online at www.phytojournal.com

J P P Journal of Pharmacognosy and Phytochemistry

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2019; 8(3): 3537-3544
Received: 04-03-2019
Accepted: 06-04-2019

V Radhika
Division of Plant Genetic Resources, ICAR-Indian Institute of Horticultural Research, Bangalore, Karnataka, India

Correspondence
V Radhika
Division of Plant Genetic Resources, ICAR-Indian Institute of Horticultural Research, Bangalore, Karnataka, India

Prediction of heat shock proteins in plants based on amino acid composition and machine learning methods

V Radhika

Abstract
Heat shock proteins (HSPs) are an important class of proteins which are expressed in cells during extreme biotic or abiotic stress conditions. Rapid identification of the HSPs is crucial in studies related to inducing plant tolerance to abiotic stresses using biotechnological approaches. In the present study we have presented a discrete model based on features of protein sequences namely sequence length along with (i) amino acid compositions (ii) di-peptide compositions and (iii) in combination and machine learning based classifiers viz. decision trees, nearest neighbour and Naïve Bayes for the identification of the heat shock proteins. A classifier for the classification of each class of heat shock proteins (HSP70, HSP90, HSP100 and sHSP) from the remaining sequences has been able developed. Based on the AUC measure, the Naïve Bayes algorithm has been found to be superior in identifying the heat shock proteins in all the classes.

Keywords: Heat shock proteins, classification, nearest neighbor, Naïve Bayes, decision tree

Introduction
Heat shock proteins, also known as stress proteins, are an important class of functionally related proteins which are expressed in the cells when they are exposed to conditions of stress like low/ high temperatures and help the organism to survive (Parsell and Lindquist, 1993) [30]. They are produced in all living organisms in response to the stress conditions. Several HSPs are known to function as "molecular chaperones," preventing aggregation and promoting the proper refolding of denatured proteins (Bukau *et al.* 2006) [5]. Although referred to as heat shock proteins, most of these proteins in fact are expressed at rather significant levels in all cells maintained under normal growth conditions and are essential for cellular growth at all physiologically relevant temperatures (Georgopoulos and Welch, 1993) [15]. Release of HSPs in plant cells following exposure to different stresses in plants has the character of an emergency response, being extremely rapid and very strong and they are induced at different induction temperatures being specific to the stress conditions for the organism (Parsell and Lindquist, 1993) [30]. Organisms induce HSP synthesis when their temperature increases above that which is specific for them, rather than at a universal temperature threshold. The induction of HSPs correlates with the induction of tolerance to extreme heat in a wide variety of cells and organisms (Li and Laszlo, 1985) [21]. The HSPs help the cells to cope up from damage to polypeptides in two ways – firstly by promoting degradation of abnormal proteins and secondly by reactivating stress-damaged proteins. In the plant system they are broadly divided into 5 families – HSP60, HSP70, HSP90, HSP100 and sHSP (small HSPs). Members of the HSP70 and HSP60 families for example, participate in protein folding, protein translocation, and perhaps higher ordered protein assembly while other members of the heat shock protein family, such as HSP90, play important roles in the regulation of certain transcription factors and protein kinases (Georgopoulos and Welch, 1993) [15]. The features that usher some stress-damaged proteins along the degradation pathway and others along the renaturation pathways are not currently understood (Parsell and Lindquist, 1993; Bukau and Horwich, 1998) [30, 4]. Researchers are investigating the role of these proteins in conferring stress tolerance to hybridized plants which can help in combating drought and poor soil conditions.
Identification of these HSPs is crucial in studies related to inducing plant tolerance to abiotic stresses using biotechnological approaches. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more known proteins (Whisstock, 2003) [40]. To get the desired results, the popular sequence- similarity-search-based tools, such as BLAST (Altschul, 1997; Wootton and Federhen, 1993) [1, 42] are usually utilized to conduct the prediction.

Other methods include identification of conserved patterns in members of a functionally uncharacterized family for which many sequences and structures are known (Whisstock, 2003) [40]. However, this kind of approach failed to work when the query protein does not have significant sequence similarity to any attribute-known proteins. Thus, various non-sequential models, or discrete models, were proposed by different workers. The simplest discrete model used to represent a protein sample is its amino acid composition (AAC) and the di-peptide compositions (DPC) (Nakashima *et al*., 1986; Chou, 2001) [28, 8]. In the present study, utility of the features of protein sequences namely length along with (i) amino acid compositions (ii) di-peptide compositions and (iii) in combination, with the help of three machine learning based classifiers namely, decision tree, nearest neighbour and Naïve Bayes have been evaluated for the identification of the heat shock proteins so as to come out with best approach.

Materials and Methods

The primary sequences of non-redundant protein sequences have been utilized for the study. Important features have been derived from the protein sequences and have been utilized for development of classification model.

Source of data

The primary sequences of the proteins have been utilized for the study. The protein sequences have been downloaded from the SwissProt (Berman *et al*., 2000; Boeckmann *et al*., 2003) [2, 3] protein database of NCBI (NCBI, 2014) [29] which contains non-redundant sequences. The protein sequences were downloaded in fasta format. The protein sequences not belonging to any of the groups- HSP60, HSP70, HSP90, HSP100, sHSP have been termed as non-HSP.

Input data representation

The sequences were converted from fasta to tabular format and stored in Excel files for further analysis. Identical sequences were identified and only one copy of the sequence was retained. A protein sequence is a chain of amino acids, which are 20 in number. The following features of all the proteins have been computed to be used as input features for the development of classification models – length, AAC (20) and DPC (400). A protein sequence denoted by 'P' of length 'N' can be represented as a sequence $X_1X_2 \ldots X_N$, where X_1, X_2,…., X_N are the amino acids. AAC and DPC for the amino acids and di-peptides in 'P' have been computed using the following formulae:

$$\text{Amino acid composition of } P_i = \frac{\text{Number of occurences of } P_i \text{ in P}}{N}$$

$$\text{Di} - \text{peptide composition of } P_iP_j = \frac{\text{Number of occurences of } P_iP_j \text{ in P}}{N-1}$$

$$for, 1 \leq i,j \leq 20$$

Matlab (Guide, 1998) [16] scripts were developed for the computation of AAC and DPC frequencies. Three different combinations of the features (Len + AAC, Len + DPC, Len + AAC + DPC) were used as input features for the development of classification models for the identification of important classes of plant heat shock protein sequences (HSP70, HSP90, HSP100 and sHSP).

Generation of Training and testing sets

10-fold cross-validation method has been followed for generating the training and testing sets for fitting and evaluation of classification models. Each dataset was randomly partitioned into 10 subsets of approximately equal size so that each class of protein is represented equally in each subset. 9 parts were used as the training set for fitting the classification model and the remaining part was used for testing the model. This process was performed over all the 10 possible combinations of training and testing datasets and the classifier efficiency parameters were evaluated for each set.

Model fitting

The classifiers namely J48 decision tree, Naïve Bayes and IB1 nearest neighbor models were fitted to the 10 sets of training and testing sets obtained in the previous step for development of the classification models. The parameters for the various models are as given in Table 1.

Table 1: Parameter values of the models J48 and nearest neighbor implemented in the study

Model	Parameter	Value
J48	Confidence threshold for pruning	0.25
	Minimum instance for each leaf	2
IB1	Distance measure	Normalized Euclidean distance

Performance evaluation

The classification models obtained based on J48 decision tree, Naïve Bayes and nearest neighbor were compared based on the performance measures. For each classification algorithm, the performance of the classifier was obtained based on the measures of accuracy, precision, recall (sensitivity), F-measure and area under the receiver operating characteristic (ROC) curve (AUC). Accuracy is the proportion of the proteins which have been classified accurately to its respective class by the classification model. Precision is the fraction of retrieved instances that are relevant while recall is the fraction of relevant instances that are retrieved. F-measure is the harmonic mean of precision and recall. The above parameters have been calculated based on the formulae:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision (p)} = \frac{TP}{TP + FP}$$

$$\text{Recall (r)} = \frac{TP}{TP + FN}$$

$$\text{F} - \text{measure} = \frac{2pr}{p + r}$$

where, TP: number of true positives, TN: number of true negatives, FP: Number of false positives and FN: number of false negatives.

Results

Frequency of HSP sequences

The sequences of different type of HSPs available in SwissProt have been tabulated (Table 2). The frequency of HSP70 was highest while that HSP60 was least among all HSPs. The number of proteins was highest in Arabidopsis

thaliana and Oryza sativa in the HSP70, HSP90 and HSP100 families while the reverse trend was observed in sHSP family.

Table 2: Frequency of the different type of HSP sequences present in SwissProt database

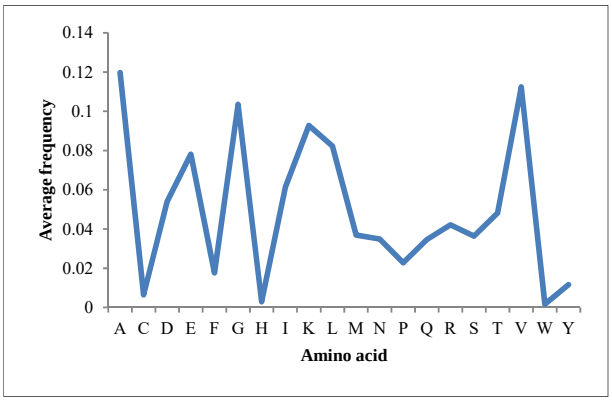
Protein type	# available in different plant species	Total
HSP60	Arabidopsis thaliana (1), Solanum tuberosum (1), Zea mays (1)	3
HSP70	Arabidopsis thaliana (103), Oryza sativa Japonica group (27), Oryza sativa Indica group (12), Solanum tuberosum (16), Nicotiana tabacum (14), etc.	226
HSP90	Arabidopsis thaliana (52), Oryza sativa Japonica group (23), Oryza sativa Indica group (11), Pisum sativum (4), Glycine max (4), etc.	139
HSP100	Arabidopsis thaliana (7), Oryza sativa Japonica group (6), Pisum sativum (1)	14
sHSP	Arabidopsis thaliana (1), Oryza sativa Japonica Group (23)	24

Average amino acid composition in different protein types

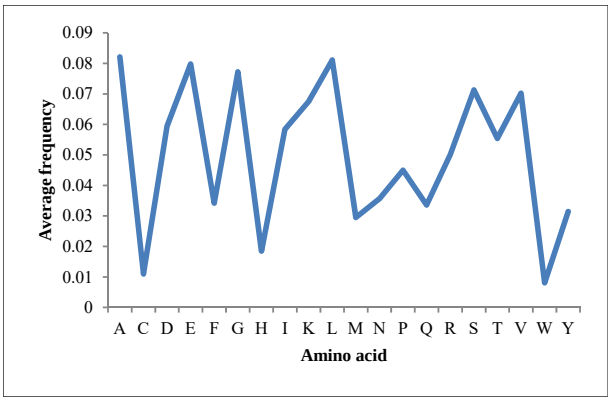
The average amino acid compositions were computed for the proteins of the HSP families as well as the non-HSPs (Table 3). Further distribution of amino acid frequencies in different classes of HSP were also depicted (Fig 1 (a-f)). In HSP60 and HSP70 families the frequency of Alanine, was highest while that of Tryptophan was lowest. In HSP90 and HSP100 families frequency of Leucine was highest while that of Tryptophan was lowest. In the sHSP family, frequency of Arginine and Glutamic acid was found to be highest while that of Cysteine was lowest. Thus a trend appeared to be present in distribution of AA in HSPs especially presence of lowest concentration of tryptophan. Tangchum *et al.* [16] showed that the three richest amino acids in HSP70 of all origins like heated cultured human leukemia cancer cell line K562, rabbit liver, rat liver and heart, and mouse liver were Glycine, Glutamic acid and Aspartic acid, except that of rat heart which was rich in Glycine, Phenylalanine and Glutamic acid. Additionally, Lysine, Valine, Leucine and Alanine were also found very rich in HSP70.

Table 3: Average amino acid composition in the different protein types

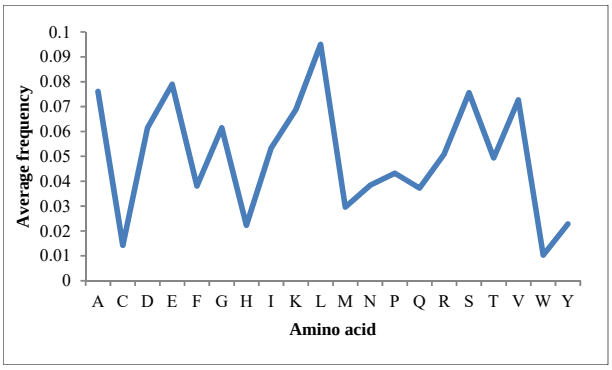
Amino acid	Symbol	Protein type					
		HSP60	HSP70	HSP90	HSP100	sHSP	Others
Alanine	A	0.120	0.082	0.076	0.085	0.108	0.073
Cysteine	C	0.006	0.011	0.014	0.007	0.005	0.019
Aspartic acid	D	0.054	0.059	0.062	0.055	0.070	0.048
Glutamic acid	E	0.078	0.080	0.079	0.086	0.086	0.059
Phenylalanine	F	0.018	0.034	0.038	0.031	0.038	0.046
Glycine	G	0.104	0.077	0.062	0.077	0.083	0.071
Histidine	H	0.003	0.019	0.022	0.015	0.017	0.022
Isoleucine	I	0.061	0.058	0.053	0.061	0.027	0.060
Lysine	K	0.093	0.068	0.069	0.063	0.061	0.060
Leucine	L	0.082	0.081	0.095	0.101	0.063	0.094
Methionine	M	0.037	0.030	0.030	0.024	0.024	0.025
Asparagine	N	0.035	0.036	0.038	0.030	0.023	0.042
Proline	P	0.023	0.045	0.043	0.040	0.064	0.048
Glutamine	Q	0.035	0.034	0.037	0.038	0.021	0.035
Arginine	R	0.042	0.050	0.051	0.075	0.086	0.055
Serine	S	0.036	0.071	0.076	0.071	0.061	0.078
Threonine	T	0.048	0.055	0.049	0.048	0.038	0.053
Valine	V	0.112	0.070	0.073	0.071	0.101	0.068
Tryptophan	W	0.002	0.008	0.010	0.004	0.016	0.013
Tyrosine	Y	0.012	0.031	0.023	0.020	0.008	0.031



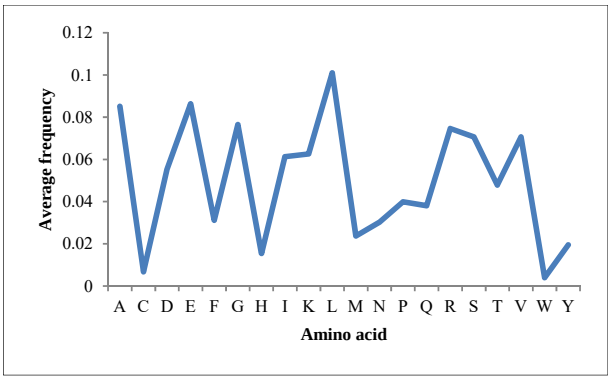
(a) HSP60



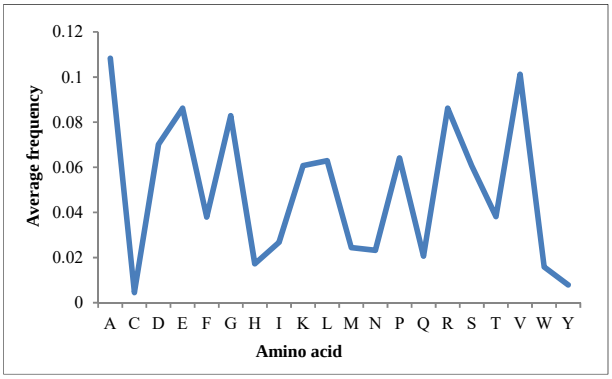
(b) HSP70



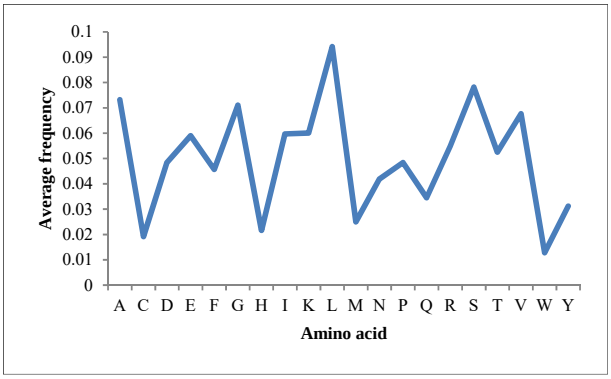
(c) HSP90



(d) HSP100



(e) sHSP



(f) Other (other than HSP60, HSP70, HSP90, HSP100 and sHSP)

Fig 1(a-f): Distribution of average amino acid frequencies in different classes of HSP proteins

Performance of the classification models

The performance metrics were obtained for all the 10 sets in all the four models and the average performance metrics of the classifiers over all the ten sets was computed for all the feature sets (Table 4 (a-d)). In case of model for identification of HSP70 sequences, the value of AUC is higher in the case of Naïve Bayes classification algorithm (0.865, 0.873 and 0.874) compared to the case of the other two classifiers for all the three cases considering different feature sets. In the case of HSP90 also the AUC values were higher in the case of Naïve Bayes in all the three data sets (0.885, 0.888 and 0.885). Similar trend was observed in HSP100 and sHSP, with the Naïve Bayes classifier having highest AUC values (0.951, 0.892 and 0.928 in HSP100; 0.999, 0.793 and 0.797 in sHSP).

Table 4(a-d): Performance metrics of the classification models for identification of different classes of HSP sequences

a. HSP70

Attributes	Model	Performance metrics*					
		Sensitivity (HSP70)	Sensitivity (others)	Accuracy	False Positive Rate	F measure	AUC
AAC	IB1	0.617	0.998	0.995	0.381	0.995	0.807
	J48	0.299	0.999	0.995	0.697	0.994	0.73
	Naïve Bayes	0.654	0.892	0.89	0.344	0.936	0.865
DPC	IB1	0.72	0.997	0.995	0.279	0.996	0.858
	J48	0.397	0.999	0.995	0.599	0.994	0.746
	Naïve Bayes	0.864	0.717	0.718	0.136	0.83	0.873
AAC + DPC	IB1	0.734	0.997	0.996	0.265	0.996	0.866
	J48	0.383	0.998	0.994	0.613	0.994	0.734
	Naïve Bayes	0.86	0.721	0.722	0.141	0.832	0.874

b. HSP90

Attributes	Model	Performance metrics*					
		Sensitivity (HSP90)	Sensitivity (others)	Accuracy	False Positive Rate	F measure	ROC
AAC	IB1	0.644	0.998	0.997	0.354	0.997	0.821
	J48	0.341	0.999	0.997	0.657	0.996	0.815
	Naïve Bayes	0.778	0.896	0.895	0.222	0.941	0.885
DPC	IB1	0.748	0.994	0.993	0.251	0.994	0.871
	J48	0.43	0.999	0.997	0.568	0.997	0.776
	Naïve Bayes	0.867	0.721	0.722	0.134	0.835	0.888
AAC + DPC	IB1	0.741	0.995	0.997	0.354	0.997	0.821
	J48	0.437	0.999	0.997	0.657	0.996	0.815
	Naïve Bayes	0.867	0.73	0.895	0.222	0.941	0.885

c. HSP100

Attributes	Model	Sensitivity (HSP100)	Sensitivity (others)	Accuracy	False Positive Rate	F measure	ROC
AAC	IB1	0.571	1	1	0.429	1	0.786
	J48	0.286	1	0.999	0.714	0.999	0.685
	Naïve Bayes	0.857	0.995	0.995	0.143	0.997	0.951
DPC	IB1	0.714	1	0.999	0.286	0.999	0.857
	J48	0.214	1	0.999	0.785	0.657	0.657
	Naïve Bayes	0.714	0.999	0.998	0.286	0.999	0.892
AAC + DPC	IB1	0.714	1	0.999	0.286	1	0.857
	J48	0.143	1	0.999	0.857	0.999	0.655
	Naïve Bayes	0.786	1	0.999	0.214	1	0.928

d. sHSP

Attributes	Model	Sensitivity (sHSP)	Sensitivity (others)	Accuracy	False Positive Rate	F measure	AUC
AAC	IB1	0.304	1	0.999	0.696	0.999	0.652
	J48	0.217	1	0.999	0.782	0.999	0.628
	Naïve Bayes	0.87	0.998	0.998	0.13	0.999	0.999
DPC	IB1	0.391	1	0.999	0.608	0.999	0.696
	J48	0.174	1	0.999	0.826	0.999	0.612
	Naïve Bayes	0.522	0.981	0.98	0.478	0.989	0.793
AAC + DPC	IB1	0.348	1	0.999	0.286	1	0.857
	J48	0.261	1	0.999	0.739	0.999	0.575
	Naïve Bayes	0.609	0.986	0.986	0.391	0.992	0.797

Comparison of the performance of the Naïve Bayes classifier with IB1 and J48

Wilcoxon matched pairs rank sum test (Wilcoxon, 1945) [41] was used to rank the different classifiers Naïve Bayes, IB1 and J48 (Table 5). Very interesting results were obtained. For the first feature set (Len, AAC) Naïve Bayes classifier got the highest rank compared to the others across all the four groups of HSPs. However, DPC resulted in similar rank for Naïve Bayes and IB1 for HSP 70 and HSP 100. Similarly, AAC+DPC resulted in similar rank for Naïve Bayes and IB1 for HSP 70 and HSP 100.

Table 5: Comparison of the Naïve Bayes classifier with IB1 and J48 using Wilcoxon matched-pair s rank sum test

HSP	Method	AAC		DPC		AAC + DPC	
		Rank sum (+, -)	p-value	Rank sum (+, -)	p-value	Rank sum (+, -)	p-value
HSP70	IB1	(55,0)	0.002	(33, 12)	0.063*	(27.5, 8.5)	0.222*
	J48	(55,0)	0.002	(55, 0)	0.002	(55, 0)	0.002
HSP90	IB1	(55,0)	0.002	(36, 0)	0.008	(36, 0)	0.008
	J48	(55,0)	0.002	(55, 0)	0.002	(55, 0)	0.002
HSP100	IB1	(55,0)	0.002	(21, 7)	0.297*	(27, 1)	0.031*
	J48	(55,0)	0.002	(55, 0)	0.002	(55, 0)	0.002
sHSP	IB1	(55,0)	0.002	(55, 0)	0.002	(55, 0)	0.002
	J48	(55,0)	0.002	(45, 0)	0.004	(55, 0)	0.002

Discussion

Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more known proteins while other methods include identification of conserved patterns in members of a functionally uncharacterized family for which many sequences and structures are known (Whisstock *et al*., 2003) [40]. However, these approaches failed to work when the query protein does not have significant sequence similarity to any attribute-known proteins. For instance, using publicly available gene expression data and predicted secondary structures, Waters *et al*. (2008) [39] have found that the sHSPs are a dynamic protein family in angiosperms which are far more diverse in sequence, expression profile, and in structure than had been previously known.

Thus, various non-sequential models, or discrete models, were proposed by different workers. The simplest discrete model used to represent a protein sample is its amino acid (AAC) composition or the di-peptide compositions (Nakashima *et al*., 1986; Chou, 2001) [28, 8].

In our study we have used the features of protein sequences namely the amino acid compositions and developed machine learning models for the identification of the proteins. Numerous studies have been reported for the classification/ identification of different types of proteins. Models has been developed for the prediction of cyclin proteins (Mohabatkar, 2010) [25], secretory proteins (Garg and Raghava, 2008) [13], protein functional class (King *et al*., 2000) [19], RNA binding sites (Kumar *et al*., 2008) [20], DNA binding proteins (Zhao *et al*, 2012) [43], subcellular localization of proteins (Garg *et al*., 2005; Mooney *et al*., 2011; Lin *et al*., 2008) [12, 26, 23], protein structural class (Cai *et al*., 2001; Lin and Li, 2007) [6, 22], the seed storage class of seed storage proteins (Radhika and Rao, 2015) [31], enzyme function (Syed and Yona, 2009) [35], protein-coding from non-coding RNAs (Liu *et al*., 2006) [24], protein enzymatic class (Volpato *et al*., 2013) [37]. So far, no study has been done on prediction of HSPs based on machine learning methods. However, few databases have been developed for HSP60 family (Hill *et al*, 2004) [18] and HSPs in general (Nagarajan *et al*, 2012) [27]. In this paper we have attempted to develop classification models for the identification of heat shock proteins viz. HSP70, HSP90, HSP100 and sHSPs.

Tangchun *et al*. (1998) [36] showed that the three richest amino acids in HSP70 of all origins like heated cultured human leukemia cancer cell line K562, rabbit liver, rat liver and heart, and mouse liver were Glycine, Glutamic acid and Aspartic acid, except that of rat heart which was rich in Glycine, Phenyl alanine and Glutamic acid. Additionally, Lysine, Valine, Leucine and Alanine were also found very rich in HSP70. Thus, there appeared to be variations in amino acid composition of HSPs of different origins. HSP70 is one of the most abundant and best characterized HSP families, expressed in response to stress and plays crucial roles in environmental stress tolerance and adaptation (Gupta *et al*. 2013) [17]. Enhanced HSP70 expression may be a response to stressful environments and may improve cell survival by protecting proteins from degradation and facilitating their refolding (Dangi *et al*. 2014) [10]. Chaurasiya *et al*. (2010) [7] found that in the four major classes of proteins namely Globins, Homeoboxes, Heat Shock proteins (HSP) and Kinase the frequency of twenty naturally occurring amino acids, hydrophobic content of protein, molecular weight of protein, isoelectric point of protein, secondary structure composition of amino acid residues as helices, coils and sheets and the composition of helices, coils and sheets in the secondary structure topology plays a significant role in correctly classifying the protein into its corresponding class or family as indicated by the overall efficiency of Nearest Neighbor Classifier as 84.92%. Sangster *et al*. (2008) [32] demonstrated that HSP90-dependent alleles occur in continuously distributed, environmentally responsive traits and are amenable to quantitative genetic mapping techniques in Arabidopsis thaliana and also found that HSP90 modulation has both general and allele-specific effects on developmental stability. However, effects of revealed variation on trait means outweigh effects of decreased developmental stability, and the HSP90-dependent trait alterations could be acted on by natural election. Thus, HSP90 may centrally influence canalization, assimilation, and the rapid evolutionary alteration of phenotype through the concealment and exposure of cryptic genetic variation.

The transcription factor ZAT12, a member of stress-responsive C2H2 type zinc finger protein (ZFP) has been reported to control the expression of stress-activated genes mediated via ROS in plants (Shah *et al*, 2013) [34]. BcZAT12-transformed tomato cv. H-86, var. Kashi vishesh (lines ZT1-ZT6) over-expressing the gene product has been demonstrated to be tolerant to heat-shock (HS)-induced oxidative stress indicating that the use of HS-tolerant tomato lines could possibly be used for tomato cultivation in the areas affected by sudden temperature changes.

There are many reports about the correlation between small molecular heat-shock protein (sHSP) and the acquirement of chilling tolerance. However, Wang at al. (2005) [38] has reported that sHSP confers enhanced chilling tolerance to plant. In their study, a DNA construct, including tomato chloroplast-localized small molecular heat-shock protein (CPsHSP) cDNA under the control of cauliflower mosaic virus 35S (35SCaMV) promoter, was introduced into the genome of tomato plants and the chilling tolerance of the transgenic tomato lines and the non-transgenic tomato was evaluated. After exposure to chilling stress, the transgenic plants exhibited lighter chilling-injured symptoms, and the results indicated consistently that transgenic tomato plants had stronger chilling tolerance. These characters are ascribed to constitutive expression of cpshsp and lead to the conclusion that HSP can enhance chilling tolerance in plant.

High-level accumulation of the target recombinant protein is a significant issue in heterologous protein expression using transgenic plants. Miraculin, a taste-modifying protein, was accumulated in transgenic tomatoes using an expression cassette in which the miraculin gene was expressed by the cauliflower mosaic virus (CaMV) 35S promoter and the heat shock protein (HSP) terminator (MIR-HSP) (Douglas at al., 2016) [41]. Further they demonstrated that the accumulation level of the target protein was comparable to levels observed with chloroplast transformation.

In our study, the dataset is highly skewed, the number of proteins belonging to the HSP classes being very low in comparison to the number of other proteins (Table 2). Of the three classifiers J48 was found to be least sensitive to the minor class (namely HSP70, HSP90, HSP100 and sHSP classes) while Naïve Bayes exhibited higher sensitivity to the minor classes. Similarly AUC values were higher in the case of Naïve Bayes across all the feature sets in all the four experiments. Very minor difference has been found in the values of AUC for the Naïve Bayes classifier for the three feature sets in the case of HSP70 and HSP90. While higher values of AUC were obtained for the Naïve Bayes classifier for the first feature set (Len, AAC) in the case of HSP100 and sHSP. Only for the first feature set (Len, AAC) Naïve Bayes classifier got the highest rank compared to the others across all the four groups of HSPs, thus it is superior. Hence, the classifier based on the feature set – length and AAC and the Naïve Bayes classifier can be recommended for the identification of heat shock proteins. This methodology can be used in conjunction with the traditional methods for the identification of heat shock proteins of the four classes as above.

Geng *et al*. (2015) [14] creatively used a 181-dimension protein sequence feature vector as input to the Naive Bayes Classifier based method to predict interaction sites in protein-protein complexes interaction. The prediction of interaction sites in protein interactions is regarded as an amino acid residue binary classification problem by applying NBC with protein sequence features. Independent test results suggested that Naive Bayes Classifier-based method with the protein sequence features as input vectors performed well. They claimed that it facilitate better understanding of biological mechanism of protein interaction which contributes to the understanding of metabolic, signal transduction networks and indicates directions in drug designing.

Douglass *et al*. (2016) [11] used a number of properties to construct the classifier, including sequence length, number of observations, existence of detectable predicted miRNA sequences, the distribution of nearby reads and mapping multiplicity for application to small RNA sequence data from soybean, peach, Arabidopsis and rice and provide experimental validation of several predictions in soybean by probabilistic method for ranking putative plant miRNAs using a naïve Bayes classifier and its publicly available implementation. The approach performed well overall and strongly enriches for known miRNAs over other types of sequences.

Cui *et al*. (2011) [9] reported a computational framework for predicting Arabidopsis mitochondrial proteins based on Naive Bayesian Network, which integrates genomic data generated from eight bioinformatics tools, multiple orthologous mappings, protein domain properties and co-expression patterns using 1,027 microarray profiles. Through this approach, they predicted 2,311 candidate mitochondrial proteins with 84.67% accuracy and 2.53% FPR performances.

Schwacke *et al*. (2007) [33] used Naïve Bayes method for *in silico*-based screening of transcription factors from Arabidopsis and rice with the aim of identifying putative N-terminal chloroplast and mitochondrial targeting sequences. In both species, transcription factors from a variety of protein families that possess putative N-terminal plastid or mitochondrial target peptides as well as nuclear localization sequences, were found.

Conclusion

In this study we have utilized few machine learning algorithms for the classification of stress protein sequences available in the public domain (NCBI). The amino acid compositions of the protein sequences have been used as input features. Classification model based on nearest neighbour algorithm has been able to classify the stress proteins more accurately. A classifier for classification of each class of heat shock proteins, HSP70, HSP90, HSP100 and sHSP) versus the remaining sequences has been able to predict the heat shock proteins. Based on the AUC performance measure, the classifiers based on Naïve Bayes method are superior in comparison to the remaining classifiers for identifying the heat shock proteins of all the classes as above.

References

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al*. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 1997; 25(17):3389-3402.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Bourne PE. The protein data bank. Nucleic Acids Research. 2000; 28(1):235-242.
3. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research. 2003; 31(1):365-370.
4. Bukau B, Horwich AL. The Hsp70 and Hsp60 chaperone machines. Cell. 1998; 92(3):351-366.
5. Bukau B, Weissman J, Horwich A. Molecular chaperones and protein quality control. Cell. 2006; 125(3):443-451.
6. Cai YD, Liu XJ, Xu XB, Zhou GP. Support vector machines for predicting protein structural class. BMC Bioinformatics. 2001; 2(1):1.
7. 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.
8. Chou KC. Prediction of protein cellular attributes using pseudo□amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001; 43(3):246-255.
9. Cui J, Liu J, Li Y, Shi T. Integrative identification of Arabidopsis mitochondrial proteome and its function exploitation through protein interaction network. PLoS One. 2011; 6(1):e16022.
10. Dangi SS, Gupta M, Nagar V, Yadav VP, Dangi SK, Shankar O, *et al*. Impact of short-term heat stress on physiological responses and expression profile of HSPs in Barbari goats. International Journal of Biometeorology. 2014; 58(10):2085-2093.
11. Douglass S, Hsu SH, Cokus S, Goldberg RB, Harada JJ, Pellegrini M. A Naive Bayesian Classifier for Identifying Plant miRNAs. The Plant Journal. 2016; 86(6):481-492.
12. Garg A, Bhasin M, Raghava GP. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. Journal of Biological Chemistry. 2005; 280(15):14427-14432.
13. Garg A, Raghava GP. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico* Biology. 2008; 8(2):129-140.
14. Geng H, Lu T, Lin X, Liu Y, Yan F. Prediction of Protein-Protein Interaction Sites Based on Naive Bayes Classifier. Biochemistry Research International, 2015.
15. Georgopoulos C, Welch WJ. Role of the major heat shock proteins as molecular chaperones. Annual Review of Cell Biology. 1993; 9(1):601-634.
16. Guide MU. The mathworks. Inc., Natick, MA, 1998; 5:333.
17. Gupta M, Kumar S, Dangi SS, Jangir BL. Physiological, biochemical and molecular responses to thermal stress in goats. International Journal of Livestock Research. 2013; 3(2):27-38.
18. Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM. cpnDB: a chaperonin sequence database. Genome Research. 2004; 14(8):1669-1675.
19. King RD, Karwath A, Clare A, Dehaspe L. Accurate prediction of protein functional class from sequence in the *Mycobacterium tuberculosis* and *Escherichia coli* genomes using data mining. Yeast. 2000; 17(4):283-293.
20. Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins: Structure, Function, and Bioinformatics. 2008; 71(1):189-194.
21. Li GC, Laszlo A. Thermotolerance in mammalian cells: A possible role for heat shock proteins. Changes in eukaryotic gene expression in response to environmental stress. 1985; 1:227-254.
22. Lin H, Li QZ. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. Journal of Computational Chemistry. 2007; 28(9):1463-1466.
23. Lin H, Ding H, Guo FB, Zhang AY, Huang J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein and Peptide letters. 2008; 15(7):739-744.
24. Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genetics. 2006; 2(4):e29.
25. Mohabatkar H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein and peptide letters. 2010; 17(10):1207-1214.
26. Mooney C, Wang YH, Pollastri G. "SCLpred: protein subcellular localization prediction by N-to-1 neural networks, Bioinformatics. 2011; 27(20):2812-2819.
27. 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.
28. Nakashima H, Nishikawa K, Tatsuo OOI. The folding type of a protein is relevant to the amino acid composition. Journal of Biochemistry. 1986; 99(1):153-162.
29. NCBI RC. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2014; 42(Database issue):D7.
30. 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.

31. Radhika V, Rao VSH. Computational approaches for the classification of seed storage proteins. Journal of Food Science and Technology. 2015; 52(7):4246-4255.
32. Sangster TA, Salathia N, Undurraga S, Milo R, Schellenberg K, Lindquist S, *et al*. HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. Proceedings of the National Academy of Sciences. 2008; 105(8):2963-2968.
33. Schwacke R, Fischer K, Ketelsen B, Krupinska K, Krause K. Comparative survey of plastid and mitochondrial targeting properties of transcription factors in Arabidopsis and rice. Molecular Genetics and Genomics. 2007; 277(6):631-646.
34. Shah K, Singh M, Rai AC. Effect of heat-shock induced oxidative stress is suppressed in BcZAT12 expressing drought tolerant tomato. Phytochemistry. 2013; 95:109-117.
35. Syed U, Yona G. Enzyme function prediction with interpretable models. Computational Systems Biology. Humana Press, 2009; 373-420.
36. Tangchun W, Yang W, Ye Y, Hanzhen H, Guogao Z. Study on amino acid composition of HSP70 and the level of plasma free amino acids of workers with long-term exposure to harmful factors. Journal of Tongji Medical University. 1998; 18(4):204-207.
37. Volpato V, Adelfio A, Pollastri G. Accurate prediction of protein enzymatic class by N-to-1 Neural Networks. BMC Bioinformatics. 2013; 14(1):1.
38. Wang L, Zhao CM, Wang YJ, Liu J. Overexpression of chloroplast-localized small molecular heat-shock protein enhances chilling tolerance in tomato plant. Journal of Plant Physiology and Molecular Biology. 2005; 31(2):167-174.
39. 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.
40. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Quarterly Reviews of Biophysics. 2003; 36(03):307-340.
41. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945; 1(6):80-83.
42. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Computers and Chemistry. 1993; 17(2):149-163.
43. Zhao XW, Li XT, Ma ZQ, Yin MH. Identify DNA-binding proteins with optimal Chou's amino acid composition. Protein and Peptide Letters. 2012; 19(4):398-405.