Send Orders for Reprints to reprints@benthamscience.ae

Letters in Organic Chemistry, **2017**, *14*, 000-000

RESEARCH ARTICLE

Predicting the Types of Plant Heat Shock Proteins

Jing Ye[1], Wei Chen[1,2,*] and Dianchuan Jin[1,2]

[1]*School of Sciences, North China University of Science and Technology, Tangshan 063000, China;* [2]*Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China*

ARTICLE HISTORY

Received: December 31, 2016
Revised: January 26, 2017
Accepted: February 03, 2017

DOI:
10.2174/1570178614666170221144023

Abstract: ***Background***: Heat shock proteins (HSPs) ubiquitously expressed in both prokaryotes and eukaryotes. According to their molecular mass and function, HSPs are classified into different families which are structurally different and play distinct functions in biological processes. Although some efforts have been made for identifying the types of HSPs, there is no method available that can be used to identify the types of HSPs in plants.

Methods: The amino acid distributions in the different types of HSPs are anazlyed. HSPs are encoded using the reduced amino acid alphabet (RAAA). By comparing the predictive capability of models based on the composition of RAAA with different sizes, the optimal feature vector was obtained. A support vector machine based model was developed to identify the types of HSPs by using the optimal feature vector.

Results: The amino acid distributions are different among the different families of HSPs. In the rigorous jackknife test, the proposed method obtained an accuracy of 93.65% for identifying the five families of HSPs in plant.

Conclusions: We hope the proposed method will become a useful tool to identify the types of HSPs in plants.

Keywords: Heat shock protein, reduced amino acid, support vector machine, n-peptide.

INTRODUCTION

Heat shock proteins (HSPs) are ubiquitously expressed in both prokaryotes and eukaryotes [1]. HSPs are stress-induced proteins that can be stimulated by physical, chemical, biological and other factors in the environment.

Researchers have demonstrated that HSPs, which behave as molecular chaperones for other cellular proteins [2], have strong cytoprotective effects and participate in many regulatory pathways. According to their molecular weight, HSPs are usually divided into six families, *i.e*. HSP20, HSP40, HSP60, HSP70, HSP90 and HSP100 [3]. These molecules are structurally different and each family plays its own functions. For example, HSP100 is responsible for regulating the activity of protein complexes [4]. As complex molecular chaperones, HSP90 is widely found in different compartments in the cell, and acts on folding newly synthesized protein and a variety of cell signal regulation [5]. HSP70 has essential functions in preventing aggregation and refolding of proteins under stress and it cooperates with HSP60 to promote cellular trafficking [6]. HSP70 can also stabilize protein by stimulating HSP70 ATPase activity together with HSP40 protein [7].

Since their discovery, HSPs have been used as models in different studies such as stress response [8] and molecular evolution. It has also been demonstrated that HSPs have different functions under different biological conditions, such as folding and unfolding of proteins, degradation of proteins, and expression of buffer mutations [9]. Although HSPs play critical roles in biological processes, their dysfunctions are associated with life-threatening diseases, including Parkinson's disease [10], cardiovascular disease [11] and cancer [12, 13]. Therefore, reliably annotating the families of HSPs is urgently important in order to clarify their functions.

HSPs not only play an irreplaceable role in plant growth and development, but also occupy a central position in translation from genotypic variation to phenotype change [14]. They were also used to treat some diseases [15]. It has also been found that functional deficiencies of HSPs can lead to morphological abnormalities and changes in physiological characteristics of plants [16]. With the characteristics of immobility, most plants have to be exposed to a variety of environmental stresses and adapt to the environment [7]. There is

*Address correspondence to this author at the Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China; Tel/Fax: +86-315-3725715;
E-mail: chenweiimu@gmail.com

1570-1786/17 $58.00+.00

evidence that HSPs are dedicated to protecting plants against stress and maintaining intracellular homeostasis [17]. In addition, HSPs are also essential for plant development [18] and maintenance of protein homeostasis [19]. HSP70 has been shown to be involved in ER quality control during seed maturation in rice [20].

Facing the rapidly increasing number of protein sequences, it is highly desired to develop automated methods for timely and reliably annotating the types of HSPs in plants. In recent years, some efforts have been made for identifying HSPs. Feng *et al.* have developed a predictor called iHSP-PseRAAAC to identify the different types of HSPs [21]. Later on, Feng *et al.* developed a freely accessible web server to classify the four types of HSP40 [22]. Stimulated by these pioneering works, Ravindra *et al.* recently established a web server called PredHSP, which is based on coupled amino acid composition and support vector machine to identify heat shock proteins and classifies their different families [23]. However, there is no method available that can be used to identify the types of HSPs in plants. Therefore, we proposed a new method to identify the types of HSPs in plants.

MATERIALS AND METHODS

Benchmark Dataset

Based on plants genome sequencing which have been completed, we have selected representative crops (*Glycine max, Hordeum vulgare, Nicotiana tabacum, Oryza sativa, Ricinus communis, Solanum lycopersicum, Sorghum bicolor, Triticum aestivum, Vitis vinifera* and *Zea mays*) and important model organisms (*Arabidopsis thaliana* and Brachypodium distachyon). The HSPs were firstlty collected from HSPIR [3] database at http://pdslab.biochem.iisc.ernet.in/hspir/. In order to enrich the dataset, based on the HSPs of *Arabidopsis thaliana*, the hmmbuild program of HMMER [24] with E-value of 1e-50 was used to search the HSPs in the remaining 11 plant proteomes. By doing so, 1275 HSPs belonging to the six HSP families (HSP20, HSP40, HSP60, HSP70, HSP90 and HSP100) were obtained.

To reduce homologous bias and redundancy, sequences which have ≥60% sequence similarity were removed by using the program CD-HIT [25] program. Finally, we obtain a benchmark dataset containing 775 sequences from five HSP families: 172 sequences belong to HSP20, 368 belong to HSP40, 100 belong to HSP60, 69 belong to HSP70 family, 66 belong to HSP100.

Representation of Protein Sequences

Based on their physiochemical properties, the 20 native amino acids can be clustered into a smaller number of representative residues knowns as reduced amino acid alphabet (RAAA) which simplifies the complexity of protein system. According to the optimization procedures as elaborated by Etchebest *et al.* [26], the 20 native amino acids can be clustered into 5 groups as shown in Table **1**. Since it was proposed, that RAAA has been increasingly used in various fields of proteomics [21, 22]. Therefore, the RAAA was used to encode the protein samples in the present study.

Table 1. Scheme for reduced amino acid alphabet.

Size	Protein Blocks Method
13	G-IV-FYW-A-L-M-E-QRK-P-ND-HS-T-C
11	G-IV-FYW-A-LM-EQRK-P-ND-HS-T-C
9	G-IV-FYW-ALM-EQRK-P-ND-HS-TC
8	G-IV-FYW-ALM-EQRK-P-ND-HSTC
5	G-IVFYW-ALMEQRK-P-NDHSTC

Accordingly, a discrete feature vector was used to represent each protein sequence as definced by

$$P = [f_1\ f_2 \cdots f_i \cdots f_D]^T \quad (1)$$

where f_i is the occurrence frequency of the *n*-peptide RAAA in protein P and T is the transposing operator. The feature vector dimension (D) of *n*-peptide (n =1, 2 or 3) composition obtained from different size (S=5, 8, 9, 11, 13) of RAAA is listed in Table **2**.

Support Vector Machine

Support vector machine (SVM) is a powerful method for pattern recognition and is widely used in the realm of bioinformatics [27-30]. The basic idea of SVM is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane. In the current study, the LibSVM package 3.18 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used to implement SVM. Due to its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the classification hyperplane in the current study. In the SVM operation engine, the grid search method was applied to optimize the regularization parameter C and kernel parameter γ using a grid search approach. The search spaces for C and γ are [2^{15}, 2^{-5}] and [2^{-5}, 2^{-15}] with the steps of 2^{-1} and 2, respectively.

Cross Validation

In statistical prediction, three cross-validation methods, *i.e.*, independent dataset test, sub-sampling (or K-fold cross-validation) test, and jackknife test, are often used to evaluate the anticipated success rate of a predictor. Among the three methods, the jackknife test is deemed as the least arbitrary and most objective [31] and has been widely used in bioinformatics [28-30]. Thus, the jackknife test was used to examine the performance of the proposed model. In the jackknife test, each sample in the training dataset is in turn singled out as an independent test sample and all the properties are calculated without including the one being identified.

Performance Evaluation

The performance of the proposed method was evaluated by using the following metrics, namely sensitivity (*Sn*), specificity (*Sp*), Accuracy (*Acc*) which are expressed as

Table 2. Feature vector dimensions of *n*-peptide composition with different RAAA sizes.

n-peptide	Dimensions of Different Amino Acid Alphabet Sizes				
	***S*=13**	***S*=11**	***S*=9**	***S*=8**	***S*=5**
n=1	13	11	9	8	5
n=2	169	121	81	64	25
n=3	2197	1331	729	512	125

Table 3. Results of HSP family classification based on different features.

Family	n-peptide Compositions of RAAA with S Size (n, S)									
	(2,13)	**(3,13)**	**(2,11)**	**(3,11)**	**(2,9)**	**(3,9)**	**(2,8)**	**(3,8)**	**(2,5)**	**(3,5)**
HSP20										
Sn(%)	93.02	98.27	92.44	96.53	86.05	98.27	84.88	97.69	73.99	85.55
Sp(%)	93.08	98.36	93.98	98.71	94.14	96.17	94.84	96.86	90.81	91.07
HSP40										
Sn(%)	96.47	98.92	96.47	98.37	94.58	98.64	93.75	97.29	84.28	91.06
Sp(%)	87.32	92.23	88.12	92.12	87.01	93.04	87.87	92.52	72.67	79.83
HSP60										
Sn(%)	67.09	83.75	67.09	80.00	67.09	75.00	72.50	76.25	62.50	61.25
Sp(%)	99.51	98.92	98.70	98.46	97.69	98.61	96.69	98.44	94.02	98.10
HSP70										
Sn(%)	57.97	72.86	59.42	77.14	65.22	71.43	63.77	72.86	47.14	60.00
Sp(%)	99.68	99.85	99.52	99.69	98.69	99.85	98.36	99.23	96.30	99.14
HSP100										
Sn(%)	78.13	86.15	79.69	83.08	79.69	83.08	76.56	81.54	47.69	64.62
Sp(%)	91.02	91.71	90.61	91.01	90.43	91.61	90.40	91.43	92.06	91.85
Acc(%)	86.92	93.65	87.05	92.73	85.34	92.21	84.81	91.55	73.05	81.51

$$\begin{cases} Sn = \dfrac{TP}{TP+FN} \times 100\% \\ Sp = \dfrac{TN}{TN+FP} \times 100\% \\ Acc = \dfrac{TP+TN}{TP+FN+TN+FP} \times 100\% \end{cases} \quad (2)$$

where *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively.

RESULTS

Amino Acids Composition Analysis

In order to analyze the distribution of the 20 native amino acids in different HSP families, the Composition Profiler [32] was used to calculate the relative amino acid enrichment or depletion in two families with the settings of *p*-value≤0.05 and Bootstrap=1000. The basic idea of Composition Profiler is the definition of fractional difference between distributions of amino acids in two different samples as follows,

$$Fractional\ difference = \frac{p_k - q_k}{q_k} \quad (3)$$

where p_k is the number of amino acid k (k represents one of the 20 native amino acids) in the query sample p and q_k is the number of amino acid k in the background sample q.

In order to analyze the amino acid distribution in a specific family of HSPs, one family of the HSPs was set as the query sample p_k, the remaining four families were used as the reference sample q_k. For instance, the remaining sequences belonging to HSP40, HSP60, HSP70, and HSP100 together form a background sample when sequences belonging to HSP20 family were used as the query sample.

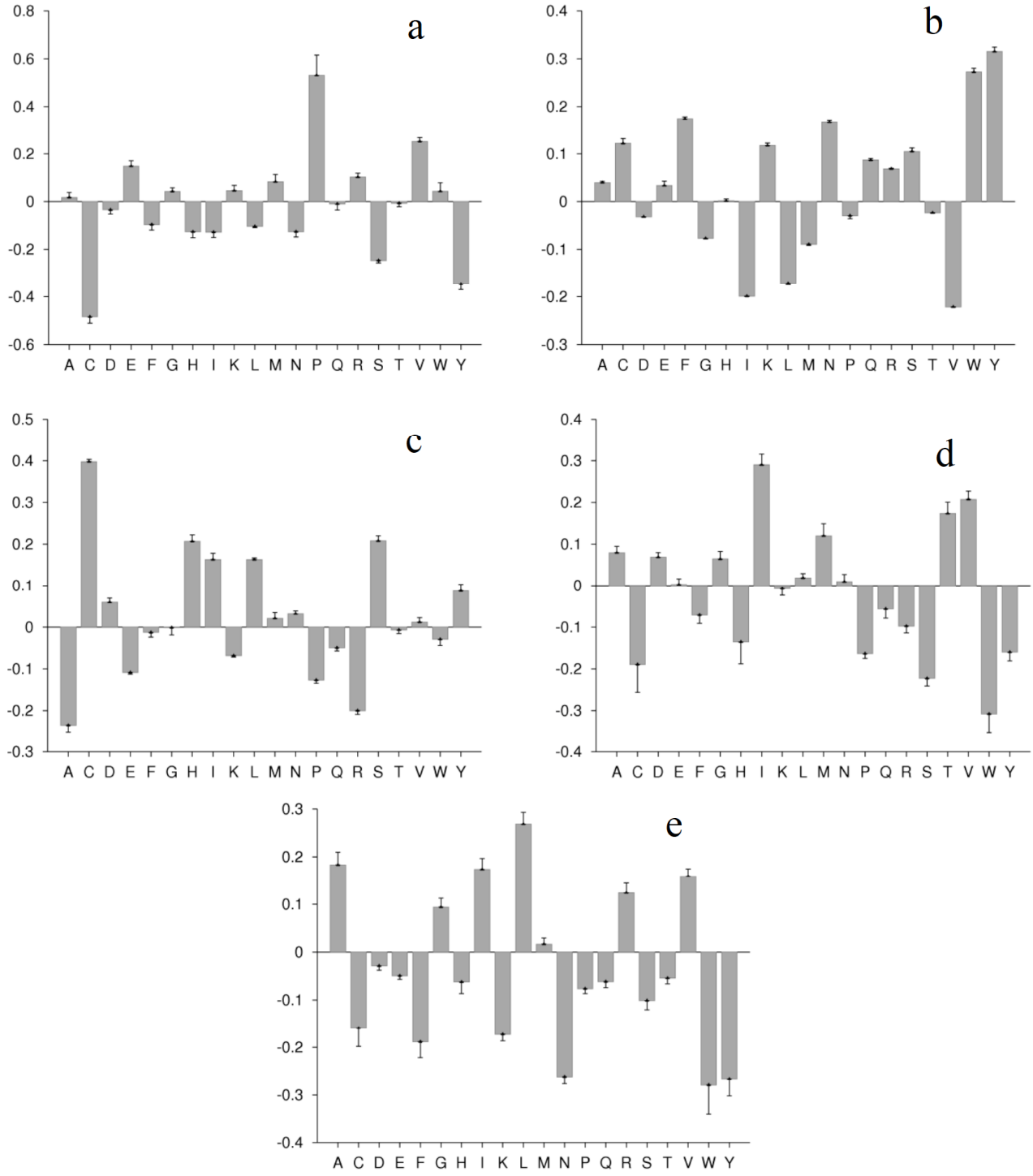

Fig. (1). An illustration to show the relative enrichment and depletion of amino acid in five HSP families, respectively. (**a**) HSP20 vs remaining HSP families; (**b**) HSP40 *vs*. remaining HSP families; (**c**) HSP60 *vs*. remaining HSP families; (**d**) HSP70 *vs*. remaining HSP families; (**e**) HSP100 vs remaining HSP families. The x-axis indicates the 20 native amino acid, and the y-axis indicates the relative abundance or depletion of the amino acids.

As shown in Figure **1**, we found that the distribution of the 20 amino acids are significantly different among the five families. Compared with the other four families, HSP20 (Fig. **1a**) are enriched in Pro (P), Val (V) but lack Cys (C), Ser (S) and Tyr (Y). Human myocytes are protected from titin aggregation-induced stiffening associated the abundance of P and H in HSP20 [33]. HSP40 (Fig. **1b**) are enriched in Trp (W) and Tyr (Y) but lack Ile (I) and Val (V). HSP60 (Fig. **1c**) are enriched in Cys (C), His (H) and Ser (S) but lack Ala (A) and Arg (R). HSP70 (Fig. **1d**) are enriched in Ile (I), Thr (T) and Val (V) but lack Cys (C), Ser (S) and Trp (W). HSP100 (Fig. **1e**) are enriched in Ala (A) and Leu (L)

but lack Phe (P), Asn (N), Trp (W) and Tyr (Y). These results indicate that the amino acid distributions are different among the five families of HSPs. Therefore, it is reasonable to develop computational tools for identifying HSP families by using sequence information.

Predicting of the Types of Hsps

In order to investigate whether a special class or property of amino acid affects the predictive accuracy or not and determine the optimal amount of information, we compared the predictive capability of models trained by using *n*-peptide (*n*=2, 3) composition of RAAA with different sizes (*S*=5, 8, 9, 11 or 13). The predictive sensitivity (*Sn*), specificity (*Sp*), matthew's correlation coefficient (*MCC*) and overall accuracy are given in Table **3**. As shown in Table **3**, the best predictive results were obtained based on 3-peptide composition of 11 reduced amino acid alphabets (*n*=3, *S*=13).

When *n*=3 and *S*=13, the overall accuracy achieved its peak, *i.e.*, Acc=93.65%, with *MCC* values of 0.82, 0.99, 0.69, 0.54, 0.30 and 0.83, respectively for HSP20, HSP40, HSP60, HSP70 and HSP100 family. These results indicate that it is a promising strategy to use RAAA to encode protein sequences and also demonstrate that the *n*-peptide composition of RAAA could extract more prominent structural and functional information than the original amino acid or dipeptide compositions.

CONCLUSION

HSPs from different families are functionally divergent. Identification of HSP family is an essential and difficult task for understanding their functions. In the present work, by collecting experimentally confirmed HSPs from 12 plants, a benchmark dataset was constructed. Based on the benchmark dataset, a support vector machine based method was developed to identify the five families of HSPs in plant. High accuracies yielded from the jackknife test indicate that the proposed method is an effective tool for HSP family identification in plant.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Program of Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

REFERENCES

[1] Kiang, J. G.; Tsokos, G. C. Heat shock protein 70 kDa: molecular biology, biochemistry, and physiology. *Pharmacol. Ther.,* **1998,** *80* (2), 183-201.

[2] Miska, K. B.; Lillehoj, H. S. Heat Shock Protein 90 Genes of Two Species of Poultry Eimeria: Expression and Evolutionary Analysis. *J. Parasitol.,* ,**2014,** *91* (2), 300-306.

[3] Ratheesh, K. R.; Nagarajan, N. S.; Arunraj, S. P.; Sinha, D.; Rajan, V. B. V.; Esthaki, V. K.; D'Silva, P. HSPIR: a manually annotated heat shock protein information resource. *Bioinformatics,* **2012,** *28* (21), 2853-2855.

[4] Burton, B. M.; Baker, T. A. Remodeling protein complexes: insights from the AAA+ unfoldase ClpX and Mu transposase. *Protein Sci.,* **2005,** *14* (8), 1945-1954.

[5] Yamada, K.; Fukao, Y.; Hayashi, M.; Fukazawa, M.; Suzuki, I.; Nishimura, M. Cytosolic HSP90 regulates the heat shock response that is responsible for heat acclimation in Arabidopsis thaliana. *J. Biol. Chem.,* **2007,** *282* (52), 37794-37804.

[6] Brocchieri, L.; Karlin, S. Conservation among HSP60 sequences in relation to structure, function, and evolution. *Protein Sci.,* **2000,** *9* (3), 476-486.

[7] Pegoraro, C.; Mertz, L. M.; Maia, L. C. D.; Rombaldi, C. V.; Oliveira, A. C. D. Importance of heat shock proteins in maize. *J. Crop Sci. Biotechnol.,* **2011,** *14* (2), 85-95.

[8] Feder, M. E.; Hofmann, G. E. Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annu. Rev. Physiol.,* **1999,** *61*, 243-282.

[9] Calderwood, S. K. Evolving connections between molecular chaperones and neuronal function. *Int. J. Hyperther.,* **2005,** *21* (5), 375-378.

[10] Aridon, P.; Geraci, F.; Turturici, G.; D'Amelio, M.; Savettieri, G.; Sconzo, G. Protective Role of Heat Shock Proteins in Parkinson's Disease. *Neurodegener. Dis.,* **2011,** *8* (4), 155-168.

[11] Madrigal-Matute, J.; Martin-Ventura, J. L.; Blanco-Colio, L. M.; Egido, J.; Michel, J. B.; Meilhac, O. Heat-shock proteins in cardiovascular disease. *Adv. Clin. Chem.,* **2011,** *54* (09), 1-43.

[12] Renaud, S.; Hajare, M.; Jessica, G.; Anne-Laure, J.; Kevin, B.; Sarah, S.; Carmen, G. Heat Shock Proteins as Danger Signals for Cancer Detection. *Front. Oncol.,* **2011,** *1*, 37.

[13] Kregel, K. C. Heat shock proteins: modifying factors in physiological stress responses and acquired thermotolerance. *J. Appl. Physiol.,* **2002,** *92* (5), 2177-2186.

[14] Sangster, T. A.; Salathia, N.; Lee, H. N.; Watanabe, E.; Schellenberg, K.; Morneau, K.; Wang, H.; Undurraga, S.; Queitsch, C.; Lindquist, S. HSP90-buffered genetic variation is common in Arabidopsis thaliana. *Proc. Nat. Acad.f Sci.,* **2008,** *105* (8), 2969-2974.

[15] Piaz, F. D.; Terracciano, S.; Tommasi, N. D.; Braca, A. Hsp90 Activity Modulation by Plant Secondary Metabolites. *Planta Medica,* **2015,** *81* (14), 1223-1239.

[16] Jackson, S. E.; Queitsch, C.; Toft, D. Hsp90: from structure to phenotype. *Nat. Struct. Mol. Biol.,* **2004,** *11* (12), 1152-1155.

[17] Shoseyov, O. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.,* **2004,** *9* (5), 244-252.

[18] Ray, D.; Ghosh, A.; Mustafi, S. B.; Raha, S. Plant Stress Response: Hsp70 in the Spotlight. **2016**.

[19] Ren, Y.; Zhu, Y. Epigenetic Regulation of Plant Heat Shock Protein (HSP) Gene Expression. Springer International Publishing: **2016**.

[20] Yuhya, W.; Hiroshi, Y.; Youko, O.; Taiji, K.; Sakiko, H.; Hideyuki, T.; Shimpei, H.; Yang, L.; Fumio, T. Expression of ER quality control-related genes in response to changes in BiP1 levels in developing rice endosperm. *Plant J.,* **2011,** *65* (5), 675-689.

[21] Feng, P. M.; Chen, W.; Lin, H.; Chou, K. C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.,* **2013,** *442* (1), 118-125.

[22] Feng, P. M.; Lin, H.; Chen, W.; Zuo, Y. Predicting the types of J-proteins using clustered amino acids. *Biomed. Res. Int.,* **2014,** *2014* (2), 935719.

[23] Kumar, R.; Kumari, B.; Kumar, M. PredHSP: Sequence Based Proteome-Wide Heat Shock Protein Prediction and Classification Tool to Unlock the Stress Biology. *Plos One,* **2016,** *11* (5).

[24] Walters, J. P.; Balu, V.; Kompalli, S.; Chaudhary, V. Evaluating the use of GPUs in liver image segmentation and HMMER database searches. *IEEE*, **2009**, 1-12.

[25] 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.

[26] Etchebest, C.; Benros, C.; Bornot, A.; Camproux, A. C.; de Brevern, A. G. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *European Biophys. J.,* **2007,** *36* (8), 1059-1069.

[27] Chen, W.; Ding, H.; Feng, P. M.; Lin, H.; Chou, K. C. iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget,* **2016,** *7* (13), 16895-16909.

[28] Chen, W.; Feng, P. M.; Tang, H.; Ding, H.; Lin, H. RAMPred: identifying the N1-methyladenosine sites in eukaryotic transcriptomes. *Scien. Rep.,* **2016,** *6*, 31080.

[29] Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K. C. iRNA-PseU: Identifying RNA pseudouridine sites. *Molecul. Ther. Nucleic Acids,* **2016**.

[30] Chen, W.; Feng, P. M.; Ding, H.; Lin, H.; Chou, K. C. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.,* **2015,** *490*, 26-33.

[31] Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor.Biol.,* **2011,** *273* (1), 236-247.

[32] Vacic, V.; Uversky, V. N.; Dunker, A. K.; Lonardi, S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics,* **2007,** *8*, 211.

[33] Kötter, S.; Unger, A.; Hamdani, N.; Lang, P.; Vorgerd, M.; Nagelsteger, L.; Linke, W. A. Human myocytes are protected from titin aggregation-induced stiffening by small heat shock proteins. *J. Cell Biol.,***2014,** *204* (2), 187-202.