Send Orders for Reprints to reprints@benthamscience.net

Current Drug Metabolism, 2019, *20*, 224-228

REVIEW ARTICLE

Recent Advances in Machine Learning Methods for Predicting Heat Shock Proteins

Wei Chen[1,3,*], Pengmian Feng[2], Tao Liu[3] and Dianchuan Jin[3]

[1]*Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China;* [2]*Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, North China University of Science and Technology, Tangshan 063000, China*; [3]*School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China*

Abstract: ***Background*:** As molecular chaperones, Heat Shock Proteins (HSPs) not only play key roles in protein folding and maintaining protein stabilities, but are also linked with multiple kinds of diseases. Therefore, HSPs have been regarded as the focus of drug design. Since HSPs from different families play distinct functions, accurately classifying the families of HSPs is the key step to clearly understand their biological functions. In contrast to labor-intensive and cost-ineffective experimental methods, computational classification of HSP families has emerged to be an alternative approach.

***Methods*:** We reviewed the paper that described the existing datasets of HSPs and the representative computational approaches developed for the identification and classification of HSPs.

***Results*:** The two benchmark datasets of HSPs, namely HSPIR and sHSPdb were introduced, which provided invaluable resources for computationally identifying HSPs. The gold standard dataset and sequence encoding schemes for building computational methods of classifying HSPs were also introduced. The three representative web-servers for identifying HSPs and their families were described.

***Conclusion*:** The existing machine learning methods for identifying the different families of HSPs indeed yielded quite encouraging results and did play a role in promoting the research on HSPs. However, the number of HSPs with known structures is very limited. Therefore, determining the structure of the HSPs is also urgent, which will be helpful in revealing their functions.

ARTICLE HISTORY

Received: January 19, 2018
Revised: May 21, 2018
Accepted: August 02, 2018

DOI:
10.2174/1389200219666181031105916

Keywords: Heat shock protein, n-peptide composition, reduced amino acid composition, machine learning, drug target, web server.

1. INTRODUCTION

Heat Shock Proteins (HSPs) are produced in cells as responses to physiological and environmental stressful conditions such as radiation, hypoxia, pH shift, nutrient deprivation, fever, cold, infection, inflammation [1, 2]. HSPs not only serve as molecular chaperones by regulating protein folding, aggregation, transport, and/or stabilization [2, 3], but also can prevent cell death by blocking the pathways [1, 4].

According to their molecular weight, HSPs can be classified into six major families [5], *i.e.* sHsp (HSP20), HSP40, HSP70, HSP60, HSP90 and HSP100. It has been demonstrated that HSPs from the six families participate in different biological processes and play distinct functions. For example, HSP40 are linked with a series of pathological conditions such as cancer, neurodegeneration, muscular dystrophy, and viral infection [6, 7]. By binding to the protein substrates, HSP70 can assist with their folding degradation, transport, and so on. HSP70 have also been reported to associate with a series of diseases, such as neurodegenerative disorders, cancer, and infectious disease [8-14]. HSP90 can regulate the conformation, stability, and activity of numerous oncogenic proteins. The inhibition of HSP90 can suppress multiple oncogenic signaling pathways [4]. Therefore, HSPs have emerged as potential drug targets and the focus of drug design.

*Address correspondence to this author at the Department of Physics, School of Sciences, North China University of Science and Technology, No.21 Bohai Road, Caofeidian Eco-city, Tangshan 063210, China; Tel/Fax: +86-315-3725715; E-mail: chenweiimu@gmail.com

With the availability of rapid sequencing technologies, a great amount of HSPs have been found. However, due to its labor-intensive nature, it is difficult to determine which families a new HSP belongs to by experimental method. Therefore, the development of computational methods for timely and reliably annotating the families of HSPs is highly desirable. Inspired by the application of machine learning methods in computational genomics and proteomics [15-21], several algorithms have been proposed for identifying HSPs.

This paper will review the existing datasets for HSPs and the representative computational approaches developed for the identification and classification of HSPs. Future perspectives of the computational prediction of HSPs are also presented.

2. BENCHMARK DATASET

2.1. Resources of Heat Shock Proteins

Heat shock protein information resource (HSPIR http://pdslab.biochem.iisc.ernet.in/hspir/) is the first curated comprehensive database for heat shock proteins [5]. HSPIR currently contains 9902 manually curated proteins which encompass 277 completed genomes covering prokayotic and eukaryotic species. These proteins in HSPIR belong to the six major families of HSP, *i.e.* sHSP (HSP20), HSP40, HSP70, HSP60, HSP90 and HSP100.

Later on, by collecting data from Uniprot [22], PFAM [23], NCBI CDD [24] and InterPro [25], an integrated resource called sHSPdb (small Heat Shock Proteins database, http://forge.info.univ-angers.fr/~gh/Shspdb/index.php) is developed for providing information about sHSP from all kingdoms [26]. At present, sHSPdb

contains approximately 4800 curated sHSP sequences and also provides a browser interface for retrieving comprehensive information on sHSPs.

2.2. Benchmark Dataset

A dataset containing many redundant samples with high similarity sequences would lack statistical representativeness. A predictor, if trained and tested by such a biased dataset, might yield misleading results with overestimated accuracy [27]. To build a high quality dataset, the CD-HIT program [28] that is widely used in computational genomics and proteomics [29-32] was used to remove HSPs with pairwise sequence identity ≥40% in HSPIR.

Accordingly, a benchmark dataset was constructed by Feng *et al*. [33], which contains 357 HSP20 sequences, 1279 HSP40 sequences, 163 HSP60 sequences, 283 HSP70 sequences, 58 HSP90 sequences and 85 HSP100 sequences. This benchmark dataset has been used to train computational models for classifying the six major families of HSPs [33].

3. SEQUENCE ENCODING SCHEMES

To develop a sequence-based predictor for identifying the attribute of a protein, one of the keys is to formulate its sequence with an effective discrete expression that can truly reflect the intrinsic correlation with the attribute to be predicted.

3.1 n-peptide Composition

The most straightforward method to formulate a protein is the *n*-peptide Composition (NPC). By doing so, a protein sequence can be converted into the following discrete vector,

$$P_{NPC} = [f_1, \cdots f_i, \cdots f_{20^n}]^T \quad (1)$$

where f_i is the occurrence frequency of the *i*-th *n*-peptide in a protein sequence and is defined as following,

$$f_i = \frac{N_i}{L-n+1} \quad (2)$$

N_i is the number of the *i*-th *n*-peptide in the protein sequence and *L* is the sequence length. When *n*=1, NPC indicates the amino acid composition; *n*=2 indicates the dipeptide composition, and so forth. The sequence encoding scheme of *n*-peptide composition has been successfully and widely used in many bioinformatics studies on peptides and proteins [21, 34-45].

3.2. Reduced Amino Acid Alphabet

Although the *n*-peptide composition can incorporate some sort of sequence order information, the dimension formed in this way will increase rapidly. For instance, the vector formed by the tripeptide composition would be 20^3=8000 dimensions, and that formed by the *n*-peptide composition would be 20^n dimensions. The high-dimension disaster problem will appear with the increase of *n*.

To alleviate such a problem, the reduced amino acid alphabet (RAAA) has been introduced to encode protein sequences [46]. In addition, by using RAAA to encode protein sequences, we could also improve the ability to find structurally conserved regions and structural similarity of entire proteins [47-49]. One common way to design RAAA is by clustering amino acids into groups according to sequence or structure information [50]. Recently, Etchebest and his colleagues [51] defined the RAAA based on a structural alphabet called Protein Blocks proposed by de Brevern *et al*. [52]. According to different optimization procedures, the 20 native amino acids can form five different cluster profiles as shown in Table **1**.

Since it has been proposed, RAAA has been widely used to represent protein sequences in computational proteomics [49, 53-54]. By using RAAA, a protein sequence can be encoded by the following discrete vector:

$$P_{RAAA} = [f_1', \cdots f_i', \cdots f_D']^T \quad (3)$$

where f_i' is the occurrence frequency of the *i*-th *n*-peptide RAAA defined as:

$$f_i' = \frac{N_i'}{L-n+1} \quad (4)$$

N_i' is the number of the *i*-th *n*-peptide (generally *n*=1, 2, or 3) RAAA in the protein sequence and *L* is also the length of the protein sequence. *D* indicates the dimension of the vector and its value depends on the cluster profiles and the value of *n*, which is indicated in Table **2**.

Table 1. Scheme for reduced amino acid alphabet [33].

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

Table 2. Dimensions of the feature vector using n-peptide RAAA of different cluster profiles [33].

n-peptide	Cluster Profiles				
	CP(13)	CP(11)	CP(9)	CP(8)	CP(5)
n=1	13	11	9	8	5
n=2	169	121	81	64	25
n=3	2197	1331	729	512	125

Table 3. List of web-servers developed for identification and classification of HSPs and their website addresses.

Methods	Website Address
iHSP-PseRAAAC	http://lin-group.cn/server/iHSP-PseRAAAC
PredHSP	http://14.139.227.92/mkumar/predhsp/index.html
JPred	http://lin-group.cn/server/Jpred

4. COMPUTATIONAL MODELS FOR IDENTIFYING HEAT SHOCK PROTEINS

In the past years, several computational methods together with the corresponding web-servers have been proposed for the identification and classification of HSPs (Table **3**). Given below are brief introductions of these representative methods.

4.1. iHSP-PseRAAAC

Based on the benchmark dataset as mentioned above, Feng *et al*. developed iHSP-PseRAAAC for identifying the six major families of HSPs [33]. iHSP-PseRAAAC encodes the sequences using the dipeptide of RAAA based the cluster profile CP(11) (Table **2**). The resulting 121-dimension feature vector was then used as the input of the Support Vector Machine (SVM) classifier to make predictions. The jackknife test result demonstrates that the overall success rate achieved by iHSP-PseRAAAC is promising in identifying the six families of HSP. For the convenience of scientific community, a freely accessible online web-server for iHSP-PseRAAAC is provided at http://lin-group.cn/server/iHSP-PseRAAAC, by which researchers can easily predict which family a query HSP belongs to.

4.2. PredHSP

Inspired by iHSP-PseRAAAC, Kumar and his colleagues proposed a two-tier SVM based method called PredHSP for the prediction and classification of HSPs [55]. The 1st-tier of PredHSP predicts whether a query protein sequence is an HSP or not, and it is trained on a dataset containing 2225 HSPs reported in the work of Feng *et al*. [33] and 10000 non-HSPs obtained from UniProt/SwissProt [22]. The 2nd-tier of PredHSP is used to identify the family to which an HSP might belong, which is trained by using the same benchmark dataset as that used in the work of Feng *et al*. For both 1st-tier and 2nd-tier of PredHSP, the protein sequence is encoded using the coupled amino acid composition (*i.e*. dipeptide composition). The performance of PredHSP is a little better than that of iHSP-PseRAAAC. An online webserver for PredHSP is provided at http://14.139.227.92/mkumar/predhsp/index.html.

4.3. JPred

Based on the structural differences, HSPs can be further classified into superfamilies that play distinct molecular functions. For example, HSP40 also known as J-protein can be classified into the following four types, Type I, Type II, Type III and Type IV J-proteins. Although J-proteins are closely related to cancer properties, the four types of J-proteins play different functions [56]. For example, Type I J-protein is tumor promoting, while Type II J-protein acts as tumor suppressors [57]. Therefore, it is also necessary to classify the types of J-proteins.

By using the tri-peptide of RAAA based on the cluster profile CP(8) (Table **2**), Feng *et al*. proposed a support vector machine based model for classifying the four families of J-proteins [58]. The model is trained based on a benchmark dataset containing 1,245 J-proteins obtained from HSPIR database, which contains 63 Type I J-proteins, 53 Type II J-proteins, 1,107 Type III J-proteins, and 22 Type IV J-proteins. In the jackknife test, the proposed method obtained an accuracy of 94.06%. A user-friendly webserver was also established and could be freely accessible at http://lin-group.cn/server/Jpred.

CONCLUSION

As one kind of molecular chaperones, HSPs have been found in all living organisms from bacteria to human. They play key roles not only in assisting proper protein conformation and maintaining the overall cellular protein homeostasis, but also in diseases such as Alzheimer's disease, Parkinson's disease, and Huntingdon disease. Since the distinct functions of HPSs from different families, accurate classification of the family and superfamily of HSPs will provide vital clues in revealing their molecular functions.

It is exciting that we witnessed the progresses in the realm of HSPs. For example, some databases, computational methods as well as web-servers have been developed in this realm over the past several years. These works indeed yielded quite encouraging results and did play a role in promoting the research on HSPs. However, there are still some challenges that need to be considered in future work. For example, the accuracy for classifying the family of HSPs still needs to be improved by enlarging the benchmark dataset and by extracting new features to represent the sequences. In addition, it is known that the function of a protein is determined by its structure. However, only 294 of the 9902 HSPs in the HSPIR are with known structure. Therefore, determining the structure of the HSPs is also urgent, which will be helpful in revealing their functions.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This work was supported by the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028) and the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Seigneuric, R.; Mjahed, H.; Gobbo, J.; Joly, A.L.; Berthenet, K.; Shirley, S.; Garrido, C. Heat shock proteins as danger signals for cancer detection. *Front. Oncol.,* **2011**, *1*, 37.

[2] Hendrick, J.P.; Hartl, F.U. Molecular chaperone functions of heat-shock proteins. *Annu. Rev. Biochem.,* **1993**, *62*, 349-384.

[3] Saibil, H. Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.,* **2013**, *14*, 630-642.

[4] Banerji, U. Heat shock protein 90 as a drug target: Some like it hot. *Clin. Cancer Res.,* **2009**, *15*, 9-14.

[5] RR, K.; NS, N.; SP, A.; Sinha, D.; Veedin Rajan, V. B.; Esthaki, V.K.; D'Silva, P. HSPIR: A manually annotated heat shock protein information resource. *Bioinformatics,* **2012**, *28*, 2853-2855.

[6] Dong, C.W.; Zhang, Y.B.; Zhang, Q.Y.; Gui, J.F. Differential expression of three *Paralichthys olivaceus* Hsp40 genes in responses to virus infection and heat shock. *Fish Shellfish Immunol.,* **2006**, *21*, 146-158.

[7] Wang, Q.; Bag, J. Induction of expression and co-localization of heat shock polypeptides with the polyalanine expansion mutant of poly(A)-binding protein N1 after chemical stress. *Biochem. Biophys. Res. Commun.,* **2008**, *370*, 11-15.

[8] Pockley, A.G. Heat shock proteins, inflammation, and cardiovascular disease. *Circulation,* **2002**, *105*, 1012-1017.

[9] Wu, Y.R.; Wang, C.K.; Chen, C.M.; Hsu, Y.; Lin, S.J.; Lin, Y.Y.; Fung, H. C.; Chang, K. H.; Lee-Chen, G.J. Analysis of heat-shock protein 70 gene polymorphisms and the risk of Parkinson's disease. *Hum. Genet.,* **2004**, *114*, 236-241.

[10] Van Noort, J.M.; Bugiani, M.; Amor, S. Heat shock proteins: Old and novel roles in neurodegenerative diseases in the central nervous system. *CNS Neurol. Disord. Drug Targets*, **2017**, *16*, 244-256.

[11] Dattilo, S.; Mancuso, C.; Koverech, G.; Di Mauro, P.; Ontario, M.L.; Petralia, C.C.; Petralia, A.; Maiolino, L.; Serra, A.; Calabrese, E.J.; Calabrese, V. Heat shock proteins and hormesis in the diagnosis and treatment of neurodegenerative diseases. *Immun. Ageing,* **2015**, *12*, 20.

[12] Urbanics, R. Heat shock proteins in stroke and neurodegenerative diseases. *Curr. Opin. Investig. Drugs*, **2002**, *3*, 1718-1719.

[13] Ciocca, D.R.; Calderwood S.K. Heat shock proteins in cancer: Diagnostic, prognostic, predictive, and treatment implications. *Cell Stress Chaperones,* **2005**, *10*, 86-103.

[14] Chatterjee, S.; Burns, T.F. Targeting heat shock proteins in cancer: A promising therapeutic approach. *Int. J. Mol. Sci.*, **2017**, *18*, pii: E1978.

[15] Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.C. iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids*, **2016**, *5*, e332.

[16] Chen, W.; Tran, H.; Liang, Z.; Lin, H.; Zhang, L.Q. Identification and analysis of the N6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci Rep.*, **2015**, *5*, 13859.

[17] Chen W.; Xing P.; Zou Q. Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci. Rep.,* **2017**, *7*, 40242.

[18] Lin, H.; Chen, W.; Ding, H. AcalPred: A sequence-based tool for discriminating between acidic and alkaline enzymes. *PLoS One*, **2013**, *8*, e75726.

[19] Lin, H.; Ding, C.; Song, Q.; Yang, P.; Ding, H.; Deng K.J.; Chen, W. The prediction of protein structural class using averaged chemical shifts. *J. Biomol. Struct. Dyn.*, **2012**, *29*, 1147-1153.

[20] Lin, H.; Liu ,W.X.; He, J.; Liu, X.H.; Ding, H.; Chen, W. Predicting cancerlectins by the optimal g-gap dipeptides. *Sci. Rep.*, **2015**, *5*, 16964.

[21] Wang, X.F.; Zhang, Y.; Wang, J.M. Prediction of protein structural class based on reliefF-SVM. *Lett. Org. Chem.*, **2017**, *14*, 696-702.

[22] UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids. Res.*, **2017**, *45*, D158-D169.

[23] Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G.A.; Tate, J.; Bateman, A. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.*, **2016**, *44*, D279-285.

[24] Marchler-Bauer, A.; Bo, Y.; Han, L.; He, J.; Lanczycki, C.J.; Lu, S.; Chitsaz, F.; Derbyshire, M.K.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; Hurwitz, D.I.; Lu, F.; Marchler, G.H.; Song, J.S.; Thanki, N.; Wang, Z.; Yamashita, R.A.; Zhang, D.; Zheng, C.; Geer, L.Y.; Bryant, S.H. CDD/SPARCLE: Functional classification of proteins *via* subfamily domain architectures. *Nucleic Acids Res.*, **2017**, *45*, D200-D203.

[25] Finn, R.D.; Attwood, T.K.; Babbitt, P.C.; Bateman, A.; Bork, P.; Bridge, A.J.; Chang, H.Y.; Dosztanyi, Z.; El-Gebali, S.; Fraser, M.; Gough, J.; Haft, D.; Holliday, G.L.; Huang, H.; Huang, X.; Letunic, I.; Lopez, R.; Lu, S.; Marchler-Bauer, A.; Mi, H.; Mistry, J.; Natale, D.A.; Necci, M.; Nuka, G.; Orengo, C.A.; Park, Y.; Pesseat, S.; Piovesan, D.; Potter, S.C.; Rawlings, N.D.; Redaschi, N.; Richardson, L.; Rivoire, C.; Sangrador-Vegas, A.; Sigrist, C.; Sillitoe, I.; Smithers, B.; Squizzato, S.; Sutton, G.; Thanki, N.; Thomas, P.D.; Tosatto, S.C.; Wu, C.H.; Xenarios, I.; Yeh, L.S.; Young, S.Y.; Mitchel,l A.L. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **2017**, *45*, D190-D199.

[26] Jaspard, E.; Hunault, G. sHSPdb: A database for the analysis of small Heat Shock Proteins. *BMC Plant Biol.*, **2016**, *16*, 135.

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

[28] Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics,* **2012**, *28*, 3150-3152.

[29] Feng, P.M.; Ding, H.; Yang, H.; Chen, W.; Lin, H.; Chou, K.C. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Ther. Nucleic Acids.*, **2017**, *7*, 155-163.

[30] Chen, W.; Yang, H.; Feng, P.M.; Ding, H.; Lin, H. iDNA4mC: Identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics*, **2017**, *33*(22), 3518-3523.

[31] Chen, W.; Ding, H.; Feng P.M.; Lin, H.; Chou, K.C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget*, **2016**, *7*, 16895.

[32] Chen W.; Lin H. Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. *Comput Biol Med.*, **2012**, *42*, 504-507.

[33] 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*, 118-125.

[34] Ru, B.; Hoen, P.A.; Nie, F.; Lin, H.; Guo, F.B.; Huang, J. PhD7Faster: Predicting clones propagating faster from the Ph.D.-7 phage display peptide library. *J. Bioinform. Comput. Biol.*, **2014**, *12*, 1450005.

[35] He, B.; Kang, J.; Ru, B.; Ding, H.; Zhou, P.; Huang, J. SABinder: A web service for predicting streptavidin-binding peptides. *Biomed. Res. Int.*, **2016**, *2016*, 9175143.

[36] Li, N.; Kang, J.; Jiang, L.; He, B.; Lin, H.; Huang, J. PSBinder: A web service for predicting polystyrene surface-binding peptides. *Biomed. Res. Int.*, **2017**, *2017*, 5761517.

[37] Lin, H.; Chen, W. Prediction of thermophilic proteins using feature selection technique. *J. Microbiol. Methods*, **2011**, *84*, 67-70.

[38] Chen, W.; Lin, H. Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. *Comput. Biol. Med.*, **2012**, *42*, 504-507.

[39] Ding, H.; Deng, E.Z.; Yuan, L.F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.C. iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *Biomed. Res. Int.*, **2014**, *2014*, 286419.

[40] Ding, H.; Liang, Z.Y.; Guo, F.B.; Huang, J.; Chen, W.; Lin, H. Predicting bacteriophage proteins located in host cell with feature selection technique. *Biomed. Res. Int.*, **2016**, *71*, 156-161.

[41] Tang, H.; Zhang, C.M.; Chen, R.; Huang, P.; Duan, C.G.; Zou, P. Identification of secretory proteins of malaria parasite by feature selection technique. *Lett. Org. Chem.*, **2017**, *14*, 621-624.

[42] Feng, Y.E.; Zhao, W. Identify protein 8-class secondary structure with quadratic discriminant algorithm based on the feature combination. *Lett. Org. Chem.*, **2017**, *14*, 625-631.

[43] Feng, P.M.; Chen, W.; Lin, H. Identifying antioxidant proteins by using optimal dipeptide compositions. *Interdiscip. Sci.*, **2016**, *8*, 186-191.

[44] Feng, P.M.; Ding, H.; Chen, W.; Lin, H. Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.*, **2013**, *2013*, 530696.

[45] Feng, P.M.; Lin, H.; Chen, W. Identification of antioxidants from sequence information using naive Bayes. *Comput. Biol. Med.*, **2013**, *2013*, 567529.

[46] Mirny L.A.; Shakhnovich E.I. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. *J. Mol Biol.,* **1999**, *291*, 177-196.

[47] Zuo, Y.; Li, Y.; Chen, Y.; Li, G.; Yan, Z.; Yang, L. PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics*, **2017**, *33*, 122-124.

[48] Zuo Y.; Lv Y.; Wei Z.; Yang L.; Li G.; Fan G. iDPF-PseRAAAC: A Web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. *PLoS One,* **2015**, *10*, e0145541.

[49] Zuo, Y.C.; Li, Q.Z. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. *Peptides*, **2009**, *30*, 1788-1793.

[50] De Brevern, A.G. New assessment of a structural alphabet. *In Silico Biol.*, **2005**, *5*, 283-289.

[51] 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. *Eur. Biophys. J.*, **2007**, *36*, 1059-1069.

[52] de Brevern, A.G.; Etchebest, C.; Hazout, S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, **2000**, *41*, 271-287.

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

[54] 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*, 118-125.

[55] 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*, e0155872.

[56] Mitra, A.; Shevde, L.A.; Samant R.S. Multi-faceted role of HSP40 in cancer. *Clin. Exp. Metastasis.,* **2009**, *26*, 559-567.

[57] Sterrenberg, J.N.; Blatch, G.L.; Edkins, A.L. Human DNAJ in cancer and stem cells. *Cancer Lett.*, **2011**, *312*, 129-142.

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