Hindawi
Computational and Mathematical Methods in Medicine
Volume 2020, Article ID 8894478, 11 pages
https://doi.org/10.1155/2020/8894478

Research Article

Identifying Heat Shock Protein Families from Imbalanced Data by Using Combined Features

Xiao-Yang Jing and Feng-Min Li

College of Science, Inner Mongolia Agricultural University, Hohhot 010018, China

Correspondence should be addressed to Feng-Min Li; fmli@imau.edu.cn

Received 28 August 2020; Revised 8 September 2020; Accepted 14 September 2020; Published 23 September 2020

Academic Editor: Hui Ding

Heat shock proteins (HSPs) are ubiquitous in living organisms. HSPs are an essential component for cell growth and survival; the main function of HSPs is controlling the folding and unfolding process of proteins. According to molecular function and mass, HSPs are categorized into six different families: HSP20 (small HSPS), HSP40 (J-proteins), HSP60, HSP70, HSP90, and HSP100. In this paper, improved methods for HSP prediction are proposed—the split amino acid composition (SAAC), the dipeptide composition (DC), the conjoint triad feature (CTF), and the pseudoaverage chemical shift (PseACS) were selected to predict the HSPs with a support vector machine (SVM). In order to overcome the imbalance data classification problems, the syntactic minority oversampling technique (SMOTE) was used to balance the dataset. The overall accuracy was 99.72% with a balanced dataset in the jackknife test by using the optimized combination feature SAAC+DC+CTF+PseACS, which was 4.81% higher than the imbalanced dataset with the same combination feature. The Sn, Sp, Acc, and MCC of HSP families in our predictive model were higher than those in existing methods. This improved method may be helpful for protein function prediction.

1. Introduction

Heat shock proteins (HSPs) are ubiquitous in living organisms. They act as molecular chaperones by facilitating and maintaining proper protein structure and function [1–4]; in addition, they are involved in various cellular processes such as protein assembly, secretion, transportation, and protein degradation [5, 6]. HSPs are rapidly expressed when the cells are exposed to physiological and environmental conditions such as elevated temperature, infection, and inflammation [7, 8]. Since the HSPs were discovered in 1962 by Ritossa [9], the HSPs have been widely studied, including their involvement in cardiovascular disease, diabetes, cancer [10–14]. According to molecular function and mass, HSPs are categorized into six different families: HSP20 (small HSPS), HSP40 (J-protein), HSP60, HSP70, HSP90, and HSP100 [15]. These families of HSPs have different functions. The HSP20 family is an ATP-independent molecular chaperone. They are efficient in preventing irreversible aggregation processes by binding denatured proteins [16]. The HSP70 family is the most highly conserved among the HSP families; it is an ATP-dependent molecular chaperone that involves protein folding and remodeling [17]. HSP40 is the cochaperone of HSP70, which participates in DNA binding, protein degradation, intracellular signal transduction, exocytosis, endocytosis, viral infection, apoptosis, and heat shock sensing [18]. HSP90 is another ATP-dependent chaperone that controls protein function and activity by facilitating protein folding, binding of ligands to their receptors or targets, or the assembly of multiprotein complexes [19]. The function of the HSP100 protein is to improve the tolerance to temperature and to promote the proteolysis of specific cellular substrates and regulation of transcription [20]. Experimental determination of HSPs are time-consuming and laborious, so it is necessary to use an effective method to predict HSPs. Recently, some computational methods for predicting HSPs have been proposed in the literature. Feng et al. developed a predictor called "iHSP-RAAAC" that selected the reduced amino acid alphabet (RAAA) as a feature vector; the overall predictive accuracy was 87.42% with the jackknife test [21]. Ahmad et al. used the split amino acid composition (SAAC), the dipeptide composition (DC), and PseAAC [22, 23] to identify HSPs; the highest overall predictive accuracy was 90.7% with the jackknife test [24]. Kumar et al. predicted HSPs and non-HSPs, and the best prediction accuracy was 72.98% by using the dipeptide composition (DC) with a 5-fold cross-validation test [25]. Meher et al. used the G-Spaced Amino Acid Pair Composition (GPC) to predict HSPs; a better result was obtained with the jackknife test [26]. Chen et al. summarized the recent advances in machine learning methods for predicting HSPs [27]. Feature selection is generally essential in a classification, and the appropriate integrated feature model generally offers higher accuracy [28]. Hence, the hybrid features have been successfully used in recent studies for constructing classifiers [29, 30]. We used the hybrid features to enhance performance. In this paper, the split amino acid composition (SAAC), the dipeptide composition (DC), the conjoint triad feature (CTF), and the pseudoaverage chemical shift (PseACS) were used to predict the HSPs with the same datasets as investigated by Feng et al. Data imbalance is always considered a problem in developing efficient and reliable prediction systems; due to an imbalanced dataset, the classifier would tend towards the majority class. Here, the syntactic minority oversampling technique (SMOTE) was used to solve the problem of imbalance. The overall accuracy was 99.72% with a balanced dataset in the jackknife test by using the optimized combination feature SAAC+DC+CTF+PseACS, which was 4.81% higher than the imbalanced dataset with the same combination feature.

Table 1: The number of sequences in HSP families.

Dataset	Family	Number of HSP samples
S_1	HSP20	357
S_2	HSP40	1279
S_3	HSP60	163
S_4	HSP70	283
S_5	HSP90	58
S_6	HSP100	85
S	Overall	2225

Table 2: The number of sequences in the independent dataset.

Families	HGNC dataset	RICE dataset	
		Wang et al.	Sarkar et al.
HSP20	11	14	—
HSP40	49	—	—
HSP60	15	4	—
HSP70	17	7	24
HSP90	4	3	—
HSP100	—	3	—
Total	96	31	24

2. Material and Methods

2.1. Dataset. The benchmark dataset was generated by Feng et al. [21]; the dataset was originally taken from the HSPIR database. In order to reduce homologous bias and redundancy, the program CD-HIT [31] was used to remove those sequences that have ≥40% pairwise sequence identity. 2225 sequences were obtained from different HSP families: the subset S_1 contains 357 sequences, the subset S_2 contains 1279 sequences, the subset S_3 contains 163 sequences, the subset S_4 contains 283 sequences, the subset S_5 contains 58 sequences, and the subset S_6 contains 85 sequences (see Table 1). The dataset can be freely downloaded from http://lin-group.cn/server/iHSP-PseRAAAC. The independent datasets include two datasets: the HGNC dataset and the RICE dataset (see Table 2). The HGNC dataset [32] has 96 human HSPs, and the RICE dataset has 55 RICE HSPs, which obtained 31 HSPs from Wang et al. [33] and 24 HSPs from a single family from Sarkar et al. [34]. The independent dataset can be freely downloaded from http://cabgrid.res.in:8080/ir-hsp.

2.2. The Prediction Model Construction Overview. The prediction model process is illustrated in Figure 1. The feature parameters were extracted for the HSPs. By using various information parameters, the prediction results show that better prediction results may be obtained by combining the following four information parameters: the split amino acid composition (SAAC), the dipeptide composition (DC), the conjoint triad feature (CTF), and the pseudoaverage chemical shift (PseACS). In SAAC, the protein sequence was split into the N-terminus segment and the C-terminus segment according to the golden ratio. Among the four feature parameters, the split amino acid composition (SAAC), the dipeptide composition (DC), and the conjoint triad feature (CTF) are based on the protein sequence, while the pseudoaverage chemical shift (PseACS) is related to the protein secondary structure. Therefore, the feature parameters involved both sequence and structure information. The four feature parameters were combined, and the syntactic minority oversampling technique (SMOTE) was used to solve the problem of the imbalance dataset. The overall accuracy (OA) was 99.72% with the balanced dataset, and the result demonstrates that the proposed method is superior to the existing methods.

2.3. Feature Extraction Techniques. In order to predict the HSPs, it is very important to choose a classifier and a set of reasonable parameters. In this paper, the split amino acid composition (SAAC), the dipeptide composition (DC) [35], the conjoint triad feature (CTF), and the pseudoaverage chemical shift (PseACS) were used to predict the HSPs.

2.3.1. Split Amino Acid Composition (SAAC). Split amino acid composition (SAAC) is a feature extraction method based on AAC. In SAAC, the protein sequence is split into various segments; then, the composition of each segment is counted separately [36–39]. It is well known that the golden ratio is ubiquitous in nature. According to the golden ratio, the protein sequence is divided into the N-terminus segment and the C-terminus segment; the ratio of the N-terminus

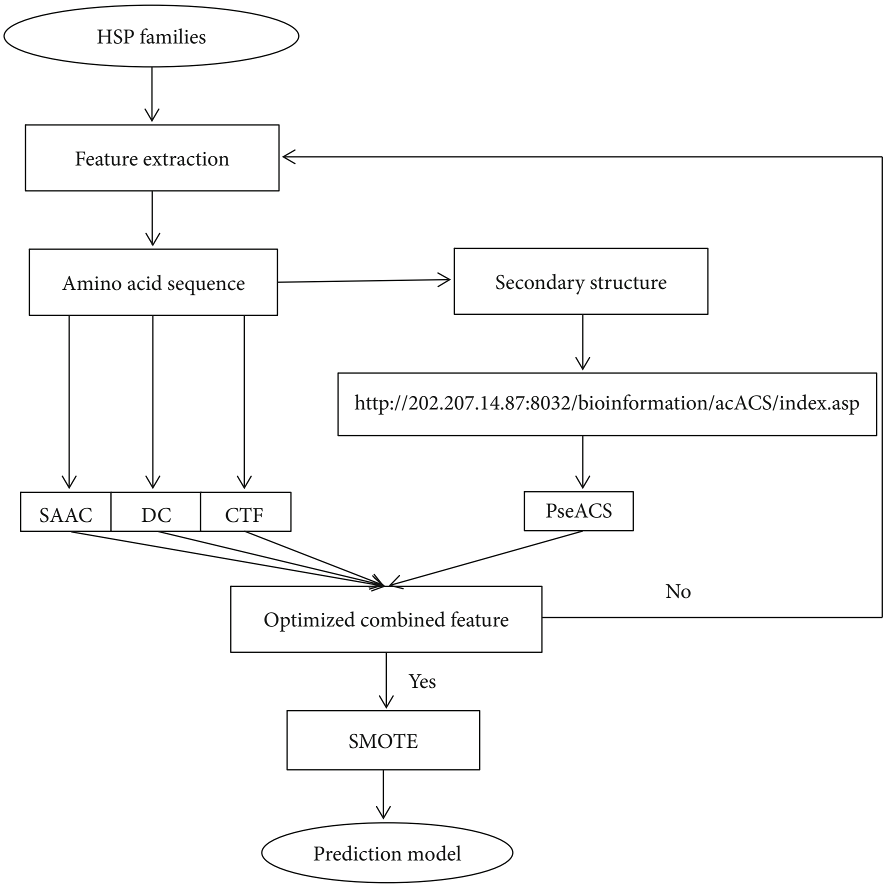

Figure 1: The flowchart of the proposed method. SAAC: split amino acid composition; DC: dipeptide composition; CTF: conjoint triad feature; PseACS: pseudoaverage chemical shift; SMOTE: syntactic minority oversampling technique.

segment to the C-terminus segment is the golden ratio [40]. This method can be represented as follows:

$$\begin{gathered}
\mathrm{SAAC\_Gr}^1 = \left(\mathrm{AAC}^{\mathrm{N}}, \mathrm{AAC}^{\mathrm{C}}\right), \\
\mathrm{AAC}^{\mathrm{N}} = \left[x_1^{\mathrm{N}}, x_2^{\mathrm{N}}, \cdots, x_i^{\mathrm{N}}, \cdots, x_{20}^{\mathrm{N}}\right], \\
\mathrm{AAC}^{\mathrm{C}} = \left[x_1^{\mathrm{C}}, x_2^{\mathrm{C}}, \cdots, x_i^{\mathrm{C}}, \cdots, x_{20}^{\mathrm{C}}\right], \\
x_i^{\mathrm{N}} = \frac{W_i}{L_{\mathrm{N}}}, \\
x_i^{\mathrm{C}} = \frac{W_i}{L_{\mathrm{C}}}, \\
(i = 1, 2, \cdots, 20),
\end{gathered} \tag{1}$$

where Gr^1 is the 1-step segmentation using the golden ratio, N represents the N-terminus, C represents the C-terminus, W_i is the occurrence of amino acid i, L_{N} is the length of the N-terminus segment, L_{C} is the length of the C-terminus segment.

With this method, we can get $\mathrm{SAAC\_Gr}^2$, $\mathrm{SAAC\_Gr}^3$, ...

$$\begin{gathered}
\mathrm{SAAC\_Gr}^2 = \left(\mathrm{AAC}_{\mathrm{N}}^{\mathrm{N}}, \mathrm{AAC}_{\mathrm{N}}^{\mathrm{C}}, \mathrm{AAC}_{\mathrm{C}}^{\mathrm{N}}, \mathrm{AAC}_{\mathrm{C}}^{\mathrm{C}}\right), \\
\mathrm{SAAC\_Gr}^3 = \left(\mathrm{AAC}_{\mathrm{NN}}^{\mathrm{N}}, \mathrm{AAC}_{\mathrm{NN}}^{\mathrm{C}}, \mathrm{AAC}_{\mathrm{NC}}^{\mathrm{N}}, \mathrm{AAC}_{\mathrm{NC}}^{\mathrm{C}}, \mathrm{AAC}_{\mathrm{CN}}^{\mathrm{N}}, \mathrm{AAC}_{\mathrm{CN}}^{\mathrm{C}}, \mathrm{AAC}_{\mathrm{CC}}^{\mathrm{N}}, \mathrm{AAC}_{\mathrm{CC}}^{\mathrm{C}}\right).
\end{gathered} \tag{2}$$

2.3.2. Dipeptide Composition (DC). Dipeptide composition (DC) is a discrete method using sequence neighbor information [27, 41, 42]. The occurrence frequency of each two adjacent amino acid residue was computed; the advantage of DC is that it considers some sequence-order information. It can be calculated as follows:

$$\begin{gathered}
P = [f_1, f_2, f_3, \cdots, f_i, \cdots, f_{400}], \\
f_i = \frac{m_i}{L - 1},
\end{gathered} \tag{3}$$

where m_i is the occurrence number of the ith dipeptide in the protein sequence, L is the length of the protein sequence.

2.3.3. Conjoint Triad Feature (CTF). The conjoint triad feature (CTF) representation was used by Shen et al. [43]. In this method, the properties of one amino acid and its vicinal amino acids were considered. Three continuous amino acids were regarded as a unit. The 20 amino acids are classified into 7 groups based on dipole moments and the volume of the side chains: $\{A, G, V\}$, $\{I, L, F, P\}$, $\{Y, M, T, S\}$, $\{H, N, Q, W\}$, $\{R, K\}$, $\{D, E\}$, and $\{C\}$. Thus, each protein sequence is represented by a 343- $(7 \times 7 \times 7)$ dimensional vector, where each element of the vector corresponds to the frequency of the corresponding conjoint triad in the protein sequence. The conjoint triad feature (CTF) has successfully predicted enzyme function [44], protein-protein interactions [45], RNA-protein interactions [46], and nuclear receptors [47]. The features of CTF can be formulated as follows:

$$\begin{aligned} \mathrm{CTF} &= [x_1, x_2, x_3, \cdots, x_i, \cdots, x_{343}], \\ x_i &= \frac{n_i}{L-2}, \end{aligned} \tag{4}$$

where n_i is the occurrence number of each triad type of the protein sequence, L is the length of the protein sequence.

2.3.4. Pseudoaverage Chemical Shift (PseACS). Nuclear magnetic resonance (NMR) plays a unique role in studying the structure of proteins because it provides information on the dynamics of the internal motion of proteins on multiple time scales [48]. Protons are sensitive to the chemical environment. The protons in different chemical environments experience slightly different magnetic fields, and they absorb different frequencies in different magnetic fields; the resonant frequencies of the various proteins in relation to a stand are called the chemical shift [49]. As important parameters are measured by nuclear magnetic resonance (NMR) spectroscopy, a chemical shift has been used as a powerful indicator of the protein structure. Several researchers revealed that the averaged chemical shift (ACS) of a particular nucleus in the protein backbone empirically correlates well to its secondary structure [50]. The PseACS web is accessible at http://202.207.14.87:8032/bioinformation/acACS/index.asp.

For a protein P, each amino acid in the sequence is substituted by its averaged chemical shift, and P can be expressed as follows:

$$P = \left[A_1^i, A_2^i, A_3^i, \cdots, A_L^i\right], \quad \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_\mathrm{N}\right), \tag{5}$$

where ^{15}N stands for nitrogen, ${}^{13}\mathrm{C}_\alpha$ for alpha carbon, ${}^{1}\mathrm{H}_\alpha$ for alpha hydrogen, and ${}^{1}\mathrm{H}_\mathrm{N}$ for hydrogen linked with nitrogen.

After, we select $\lambda = 54$ and $i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}$, the PseACS would be expressed as follows:

$$\begin{aligned} \phi_i^\lambda &= \frac{1}{L-\lambda} \sum_{k=1}^{L-\lambda} \left[A_k^i - A_{k+\lambda}^i\right]^2, \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_\mathrm{N}\ ;\ \lambda < L\right), \\ \mathrm{PseACS} &= \left[\phi_i^0, \phi_i^1, \phi_i^2, \cdots, \phi_i^\lambda\right], \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_\mathrm{N}\right). \end{aligned} \tag{6}$$

2.4. Syntactic Minority Oversampling Technique (SMOTE). As shown in Table 1, the numbers of HSP40 are about 4 times, 8 times, 5 times, 22 times, and 15 times that of HSP20, HSP60, HSP70, HSP90, and HSP100, respectively. This leads to imbalance data classification problems. In order to overcome this problem, we used the SMOTE to solve the problem of imbalance. SMOTE is an oversampling approach where the minority class is oversampled by selecting the minority class and creating new synthetic samples along the line segments connecting any or all K-Nearest Neighbors which belong to that class [51, 52]. In this paper, the protein numbers of six subfamilies are in equilibrium with SMOTE. This algorithm is implemented by the Weka software. A filter selects SMOTE when the data is loaded, and the parameters adopt the default parameters according to the number of families from small to large; the number of the remaining five families increases in turn to the number of HSP40, which is the largest number of the HSP families. In this way, SMOTE is realized.

2.5. Support Vector Machine (SVM). The support vector machine is a machine learning algorithm, which is based on the statistical learning theory. The basic idea of SVM is to transform the input data into a high-dimensional Hilbert space and then determine the optional separating hyperplane [53, 54]. The radical basis kernel function (RBF) was used to obtain the classification hyperplane with its effectiveness and speed in the training process. To handle a multiclass problem, the regulation parameter c and kernel width parameterγ were determined via the grid search method. "One-versus-one (OVO)" and "one-versus-rest (OVR)" methods are generally applied to extend the traditional SVM. In this study, the "OVO" strategy was used. The OVO strategy constructs $k \times (k-1)/2$ classifiers with each one trained with the data from two different classes. SVM has been successfully applied in the field of computational biology and bioinformatics [55–64]. In this paper, the LibSVM package was used to predict HSPs, which can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm.

2.6. Performance Evaluation. In statistical prediction, three cross-validation tests are commonly used to examine a predictor for its effectiveness in practical application: the k-fold cross-validation (subsampling test), the independent dataset test, and the jackknife test. Among the three methods, the jackknife test is deemed the most objective and rigorous one. In the jackknife test, each sample in the training dataset is in turn singled out as an independent test sample and all the rule parameters are calculated based on the remaining dataset without including the one being identified. Hence, the jackknife test was used to evaluate performance in this paper. To evaluate the predictive capability and reliability of our model, the performance of the classification algorithm is measured using the following: sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthew's correlation coefficient (MCC), and overall accuracy (OA) [65–75]. The performance of the classification algorithm is measured through the following:

$$\begin{aligned} \mathrm{Sn} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})}}, \\ \mathrm{Acc} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\ \mathrm{OA} &= \sum_{i=1}^{m} \mathrm{TP}_i/N, \end{aligned} \tag{7}$$

Table 3: The predictive results of individual features with the jackknife test by using SVM for HSP families.

Features		HSP families						OA (%)
		HSP20	HSP40	HSP60	HSP70	HSP90	HSP100	
CTF	Sn (%)	74.86	90.92	54.72	67.27	53.85	67.9	80.92
	Sp (%)	95.07	76.19	98.71	96.48	99.86	99.52	
	MCC	0.7	0.68	0.63	0.66	0.69	0.75	
	Acc (%)	91.79	84.68	95.5	92.75	98.76	98.35	
SAAC	Sn (%)	81.07	97.53	58.49	75.9	57.69	74.07	87.25
	Sp (%)	97.7	81.06	99.36	98.26	100	99.48	
	MCC	0.81	0.81	0.7	0.78	0.76	0.78	
	Acc (%)	95	90.55	96.38	95.41	98.99	98.53	
DC	Sn (%)	90.96	96.66	68.55	84.89	63.46	77.78	90.69
	Sp (%)	96.66	90.69	99.11	98.16	100	99.86	
	MCC	0.85	0.88	0.75	0.84	0.79	0.86	
	Acc (%)	95.73	94.13	96.88	96.47	99.13	99.04	
PseACS	Sn (%)	92.37	95.46	75.47	87.41	67.31	83.95	91.38
	Sp (%)	99.01	89.94	98.71	98.16	99.91	99.33	
	MCC	0.92	0.86	0.77	0.86	0.79	0.83	
	Acc (%)	97.94	93.12	97.02	96.79	99.13	98.76	

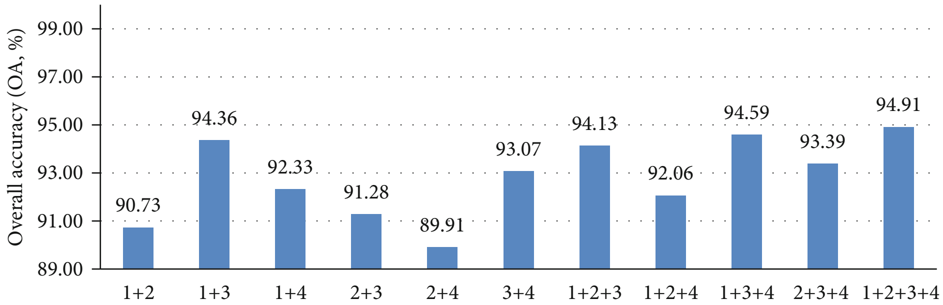

Figure 2: Prediction results of different combined features. Numbers denote features: 1 for DC, 2 for CTF, 3 for PseACS, and 4 for SAAC.

Table 4: The predictive results of HSPs by using the combined feature of SAAC+DC+CTF+PseACS with and without SMOTE.

Features with and without SMOTE (Y/N)			HSP families						OA (%)
			HSP20	HSP40	HSP60	HSP70	HSP90	HSP100	
PseACS+DC+SAAC+CTF	Y	Sn (%)	100	98.33	100	100	100	100	99.72
		Sp (%)	99.92	100	99.92	99.82	100	100	
		MCC	1	0.99	1	0.99	1	1	
		Acc (%)	99.93	99.72	99.93	99.85	100	100	
PseACS+DC+SAAC+CTF	N	Sn (%)	94.35	98.89	81.13	90.29	75	91.36	94.91
		Sp (%)	98.58	94.26	99.6	98.84	100	99.9	
		MCC	0.92	0.94	0.87	0.90	0.86	0.94	
		Acc (%)	97.89	96.93	98.26	97.75	99.4	99.59	

where TP represents the true positive, TN represents the true negative, FP represents the false positive, and FN represents the false negative. $m = 6$ is the number of subsets, and N is the number of total sequences of HSP families.

3. Results and Discussion

3.1. The Predictive Performance of HSPs. In order to investigate the effectiveness of the predictive model, many

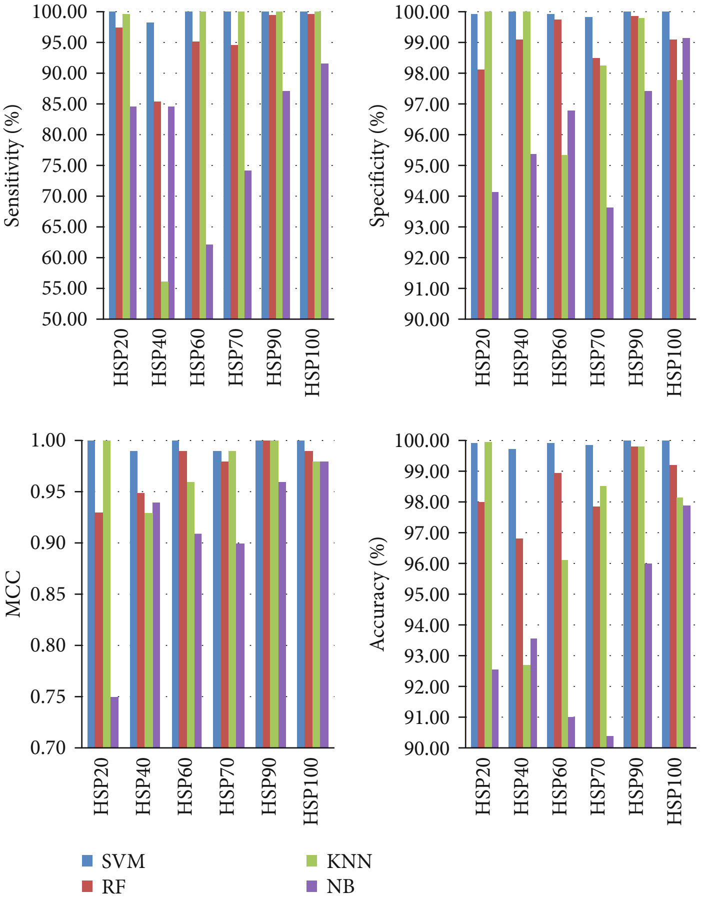

Figure 3: The predictive sensitivity, specificity, MCC, and accuracy of HSPs by using four algorithms.

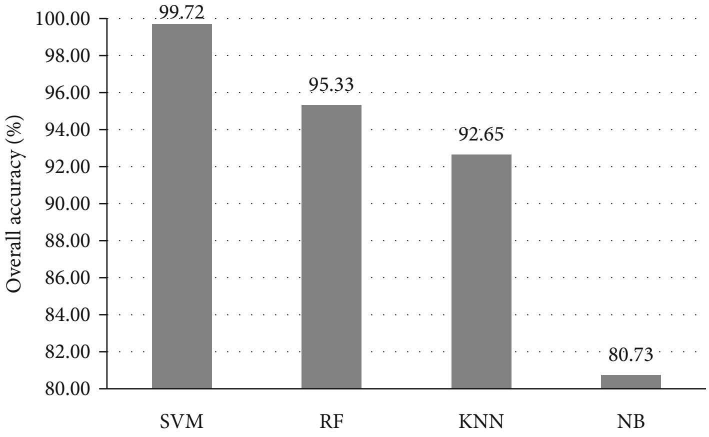

Figure 4: The predictive overall accuracy of HSPs by using four algorithms.

characteristic parameters were selected to predict the HSPs [76, 77]. Then, the split amino acid composition (SAAC), the dipeptide composition (DC), the conjoint triad feature (CTF), and the pseudoaverage chemical shift (PseACS) were selected to predict the HSPs. Table 3 lists the predictive performance of HSPs using individual features with the SVM classification algorithm without SMOTE; the highest overall accuracy (OA) of an individual parameter is 91.38% with the jackknife test by using PseACS. Individual features identify the families of HSPs with an overall accuracy (OA) ranging from 80.92% to 91.38%.

Figure 2 shows the predictive results of different combined features of HSPs with SVM without SMOTE. The results show that the combined feature of SAAC+DC+CTF

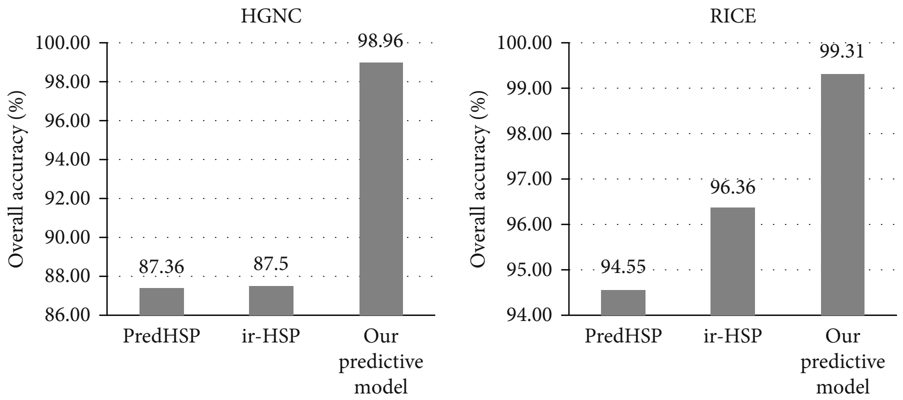

Figure 5: A comparison of the proposed method for independent datasets.

Table 5: The comparison of the predictive results between this paper and existing methods.

Method		HSP families					
		HSP20	HSP40	HSP60	HSP70	HSP90	HSP100
iHSP-PseRAAAC[a]	Sn (%)	87.68	95.31	66.87	79.15	51.72	69.41
	Sp (%)	96.36	84.87	98.93	86.54	99.89	99.84
	MCC	0.82	0.99	0.69	0.54	0.3	0.83
	Acc (%)	—	—	—	—	—	—
PredHSP[b]	Sn (%)	92.16	96.09	79.75	91.17	72.41	82.35
	Sp (%)	97.16	86.26	97.24	91.97	99.12	98.08
	MCC	0.87	0.83	0.72	0.71	0.7	0.71
	Acc (%)	96.36	91.91	95.96	91.87	98.43	97.48
ir-HSP[c]	Sn (%)	94.63	97.45	67.92	88.49	75	88.89
	Sp (%)	96.61	95.13	98.86	98.84	99.76	99.57
	MCC	0.8718	0.9276	0.7307	0.8871	0.8112	0.8846
	Acc (%)	96.28	96.47	96.61	97.52	99.17	99.17
Our predictive model	Sn (%)	100	98.33	100	100	100	100
	Sp (%)	99.92	100	99.92	99.82	100	100
	MCC	1	0.99	1	0.99	1	1
	Acc (%)	99.93	99.72	99.93	99.85	100	100

[a] Feng et al. [21]. [b] Kumar et al. [25]. [c] Meher et al. [26].

+PseACS was better than the other parameters. The overall accuracy (OA) of the combined feature of SAAC+DC+CTF +PseACS was 94.91% with the jackknife test. This result indicated that the combined feature was powerful in predicting HSPs.

Table 4 lists the predictive performance of HSP families using the optimized combination feature SAAC+DC+CTF +PseACS with and without SMOTE. In the models with SMOTE, the Sn, Sp, Acc, and MCC of HSP families improved remarkably. For example, for HSP20 with SMOTE, Sn = 100%, Sp = 99.92%, MCC = 1, and Acc = 99.93%, which are 5.65%, 1.34%, 0.08, and 2.04% higher than those without SMOTE. In addition, OA = 99.72% with SMOTE, which is 4.81% higher than HSP families without SMOTE. The results indicate that the combined parameter SAAC+DC+CTF +PseACS with SMOTE was helpful in enhancing predictive performance.

3.2. Comparison with Other Algorithms. The predictive performance of our predictive model (SVM), Random Forest (RF) [78], Naive Bayes (NB), and K-Nearest Neighbors (KNN) [79] is shown in Figures 3 and 4. From Figure 3, we can see that the differences of the Sn, Sp, MCC, and Acc of the HSP families are obvious. The Sn of HSP60, HSP70, HSP90, and HSP100 using SVM and KNN were all 100%. The Sp of HSP20 using KNN and SVM were similar, and the Sp of HSP40 using SVM and KNN were 100%. The MCC of HSP20 and HSP90 using SVM and KNN were both 1. The Acc of HSP20 using KNN and SVM were similar. In addition, from Figure 4, we can see that the value of OA with SVM was 99.72%, which was 4.39%, 7.07%, and 18.99% higher than RF, KNN, and NB, respectively. The highest value of the other parameters was obtained by SVM. Therefore, the experimental results show that SVM has achieved the best measures.

Figure 5 shows the predictive performance of HSP families using independent datasets. In the HGNC independent dataset, the OA of our predictive model was 98.96%, which was 11.60% and 11.46% higher than PredHSP and ir-HSP, respectively. In the RICE independent dataset, the OA of our predictive model reached 99.31%, which was 4.76% and 2.95% higher than PredHSP and ir-HSP, respectively. From the comparison, we can draw a conclusion that the applicability and accuracy of our prediction model for HSP prediction were improved.

3.3. Comparison with Existing Methods. In order to evaluate the performance of our predictive model, we made comparisons with existing methods. The method developed by Ahmad et al. did not provide any family-wise accuracy of HSPs, so we compared the effectiveness with iHSP-PseRAAAC, PredHSP, and ir-HSP. The results of the comparisons are shown in Table 5. We can see that the Sn, Sp, Acc, and MCC of HSP families in our predictive model were higher than those of PredHSP, iHSP-PseRAAAC, and ir-HSP. For example, in our predictive model, Sn = 100%, Sp = 99.92%, MCC = 1, and Acc = 99.93% for HSP20 exceeded those of ir-HSP, PredHSP, and iHSP-PseRAAAC. In addition, in our predictive model, Sn = 100 for all HSP families, except for HSP40 Sn = 98.33%. Furthermore, the overall accuracy was 99.72% in our predictive model. These results indicate that our predictive model was superior to existing methods.

4. Conclusion

In this work, an optimized classifier for HSP family identification was developed. This model was derived from the SVM machine learning algorithm, and SMOTE was used for the imbalanced data classification problems. The overall accuracy was 99.72% with the balanced dataset and the jackknife test by using the optimized combination feature SAAC +DC+CTF+PseACS. High overall accuracy results indicate that our predictive model is a reliable tool for HSP family prediction. It is known that HSP expression is associated with human diseases, and these families of HSPs have different functions. Therefore, our predictive model will benefit researchers by quickly and effectively identifying HSP families and enabling researchers to design new drugs to achieve the goal of treating diseases.

Data Availability

The data used to support the findings of this study are available from the supplementary materials.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Authors' Contributions

FM Li conceived the selection of feature parameters. XY Jing carried out the computation and wrote the manuscript. FM Li performed the results analysis. Both authors reviewed the manuscript.

Acknowledgments

This work was supported by the Natural Science Foundation of Inner Mongolia Autonomous Region of China (2019MS03015) and the National Natural Science Foundation of China (31360206).

Supplementary Materials

Supplementary 1. The sequence names of HSP families.

Supplementary 2. The sequence names of the independent datasets.

References

[1] T. Liu, C. K. Daniels, and S. Cao, "Comprehensive review on the HSC70 functions, interactions with related molecules and involvement in clinical diseases and therapeutic potential," *Pharmacology & Therapeutics*, vol. 136, no. 3, pp. 354–374, 2012.

[2] J. M. Wu, T. E. Liu, Z. Rios, Q. B. Mei, X. K. Lin, and S. S. Cao, "Heat shock proteins and cancer," *Trends in Pharmacological Sciences*, vol. 38, no. 3, pp. 226–256, 2017.

[3] M. E. Feder and G. E. Hofmann, "Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology," *Annual Review of Physiology*, vol. 61, no. 1, pp. 243–282, 1999.

[4] S. R. Qazi, N. Ul Haq, S. Ahmad, and S. N. Shakeel, "HSEAT: a tool for plant heat shock element analysis, motif identification and analysis," *Current Bioinformatics*, vol. 15, no. 3, pp. 196–203, 2020.

[5] S. Chatterjee and T. F. Burns, "Targeting heat shock proteins in cancer: a promising therapeutic approach," *International Journal of Molecular Sciences*, vol. 18, no. 9, p. 1978, 2017.

[6] C. Jolly and R. I. Morimoto, "Role of the heat shock response and molecular chaperones in oncogenesis and cell death," *Journal of the National Cancer Institute*, vol. 92, no. 19, pp. 1564–1572, 2000.

[7] A. Khadir, S. Kavalakatt, P. Cherian et al., "Physical exercise enhanced heat shock protein 60 expression and attenuated inflammation in the adipose tissue of human diabetic obese," *Frontiers in Endocrinology*, vol. 9, p. 16, 2018.

[8] P. C. Ikwegbue, P. Masamba, L. S. Mbatha, B. E. Oyinloye, and A. P. Kappo, "Interplay between heat shock proteins, inflammation and cancer: a potential cancer therapeutic target," *American Journal of Cancer Research*, vol. 9, no. 2, pp. 242–249, 2019.

[9] F. Ritossa, "A new puffing pattern induced by temperature shock and DNP in Drosophila," *Experientia*, vol. 18, no. 12, pp. 571–573, 1962.

[10] B. Rodríguez-Iturbe and R. J. Johnson, "Heat shock proteins and cardiovascular disease," *Physiology international*, vol. 105, no. 1, pp. 19–37, 2018.

[11] M. Zilaee and S. Shirali, "Heat shock proteins and diabetes," *Canadian Journal of Diabetes*, vol. 40, no. 6, pp. 594–602, 2016.

[12] G. D. Lianos, G. A. Alexiou, A. Mangano et al., "The role of heat shock proteins in cancer," *Cancer Letters*, vol. 360, no. 2, pp. 114–118, 2015.

[13] T. Zhao, Y. Hu, J. Peng, and L. Cheng, "DeepLGP: a novel deep learning method for prioritizing lncRNA target genes," *Bioinformatics*, 2020.

[14] C. Liang, Q. Changlu, Z. He, F. Tongze, and Z. Xue, "gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Research*, vol. 48, no. D1, pp. D554–D560, 2019.

[15] N. S. Nagarajan, S. P. Arunraj, D. Sinha, V. B. Rajan, V. K. Esthaki, and P. D'Silva, "HSPIR: a manually annotated heat shock protein information resource," *Bioinformatics*, vol. 28, no. 21, pp. 2853–2855, 2012.

[16] T. Mahmood, W. Safdar, B. H. Abbasi, and S. S. Naqvi, "An overview on the small heat shock proteins," *African Journal of Biotechnology*, vol. 9, no. 7, pp. 927–939, 2010.

[17] O. Genest, S. Wickner, and S. M. Doyle, "Hsp 90 and Hsp 70 chaperones: collaborators in protein remodeling," *The Journal of Biological Chemistry*, vol. 294, no. 6, pp. 2109–2120, 2019.

[18] T. Chen, T. H. Lin, H. M. Li et al., "Heat shock protein 40 (HSP40) in pacific white shrimp (Litopenaeus vannamei): molecular cloning, tissue distribution and ontogeny, response to temperature, acidity/alkalinity and salinity stresses, and potential role in ovarian development," *Frontiers in Physiology*, vol. 9, p. 1784, 2018.

[19] F. H. Schopf, M. M. Biebl, and J. Buchner, "The HSP90 chaperone machinery," *Nature Reviews. Molecular Cell Biology*, vol. 18, no. 6, pp. 345–360, 2017.

[20] E. C. Schirmer, J. R. Glover, M. A. Singer, and S. Lindquist, "HSP100/Clp proteins: a common mechanism explains diverse functions," *Trends in Biochemical Sciences*, vol. 21, no. 8, pp. 289–296, 1996.

[21] P. M. Feng, W. Chen, H. Lin, and K. C. Chou, "iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[22] P. F. Du, W. Zhao, Y. Y. Miao, L. Y. Wei, and L. Wang, "UltraPse: a universal and extensible software platform for representing biological sequences," *International Journal of Molecular Sciences*, vol. 18, no. 11, p. 2400, 2017.

[23] J. Wang, P. F. Du, X. Y. Xue et al., "VisFeature: a stand-alone program for visualizing and analyzing statistical features of biological sequences," *Bioinformatics*, vol. 36, no. 4, pp. 1277-1278, 2019.

[24] S. Ahmad, M. Kabir, and M. Hayat, "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*, vol. 122, no. 2, pp. 165–174, 2015.

[25] R. Kumar, B. Kumari, and M. Kumar, "PredHSP: sequence based proteome-wide heat shock protein prediction and classification tool to unlock the stress biology," *PLoS One*, vol. 11, no. 5, p. e0155872, 2016.

[26] P. K. Meher, T. K. Sahu, and S. Gahoi, "ir-HSP: improved recognition of heat shock proteins, their families and sub-types based on g-spaced di-peptide features and support vector machine," *Frontiers in Genetics*, vol. 8, p. 235, 2018.

[27] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, 2019.

[28] L. Q. Li, S. J. Yu, W. D. Xiao et al., "Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach," *Biochimie*, vol. 104, pp. 100–107, 2014.

[29] L. N. Zhang and C. J. Zhang, "JPPRED: prediction of types of J-proteins from imbalanced data using an ensemble learning method," *BioMed Research International*, vol. 2015, 12 pages, 2015.

[30] F. M. Li and X. Q. Wang, "Identifying anticancer peptides by using improved hybrid compositions," *Scientific Reports*, vol. 6, no. 1, p. 33910, 2016.

[31] W. Z. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.

[32] H. H. Kampinga, J. Hageman, M. J. Vos et al., "Guidelines for the nomenclature of the human heat shock proteins," *Cell Stress & Chaperones*, vol. 14, no. 1, pp. 105–111, 2009.

[33] Y. Wang, S. Lin, Q. Song et al., "Genome-wide identification of heat shock proteins (Hsps) and Hsp interactors in rice: Hsp70s as a case study," *BMC Genomics*, vol. 15, no. 1, pp. 344–344, 2014.

[34] N. K. Sarkar, P. Kundnani, and A. Grover, "Functional analysis of Hsp70 superfamily proteins of rice (Oryza sativa)," *Cell Stress & Chaperones*, vol. 18, no. 4, pp. 427–437, 2013.

[35] X. J. Zhu, C. Q. Feng, H. Y. Lai, W. Chen, and H. Lin, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowledge-Based Systems*, vol. 163, pp. 787–793, 2019.

[36] K. Ahmad, M. Waris, and M. Hayat, "Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition," *The Journal of Membrane Biology*, vol. 249, no. 3, pp. 293–304, 2016.

[37] M. Arif, M. Hayat, and Z. Jan, "iMem-2LSAAC: a two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 442, pp. 11–21, 2018.

[38] M. Tahir and M. Hayat, "iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC," *Molecular BioSystems*, vol. 12, no. 8, pp. 2587–2593, 2016.

[39] V. Saravanan and P. T. V. Lakshmi, "Dualpred: a webserver for predicting plant proteins dual-targeted to chloroplast and mitochondria using split protein-relatedness-measure feature," *Current Bioinformatics*, vol. 10, no. 3, pp. 323–331, 2015.

[40] Q. Dai, S. Ma, Y. B. Hai, Y. H. Yao, and X. Q. Liu, "A segmentation based model for subcellular location prediction of apoptosis protein," *Chemometrics and Intelligent Laboratory Systems*, vol. 158, pp. 146–154, 2016.

[41] W. Yang, X. J. Zhu, J. Huang, H. Ding, and H. Lin, "A brief survey of machine learning methods in protein sub-Golgi localization," *Current Bioinformatics*, vol. 14, no. 3, pp. 234–240, 2019.

[42] J. X. Tan, S. H. Li, Z. M. Zhang et al., "Identification of hormone binding proteins based on machine learning methods," *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2466–2480, 2019.

[43] J. W. Shen, J. Zhang, X. M. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[44] Y. C. Wang, Y. Wang, Z. X. Yang, and N. Y. Deng, "Support vector machine prediction of enzyme function with conjoint

triad feature and hierarchical context," *BMC Systems Biology*, vol. 5, no. S1, p. S6, 2011.

[45] J. Wang, L. Zhang, L. Jia, Y. Ren, and G. Yu, "Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences," *International Journal of Molecular Sciences*, vol. 18, no. 11, p. 2373, 2017.

[46] H. C. Wang and P. F. Wu, "Prediction of RNA-protein interactions using conjoint triad feature and chaos game representation," *Bioengineered*, vol. 9, no. 1, pp. 242–251, 2018.

[47] H. C. Wang and X. H. Hu, "Accurate prediction of nuclear receptors with conjoint triad feature," *BMC Bioinformatics*, vol. 16, no. 1, p. 402, 2015.

[48] P. Calligari and D. Abergel, "Multiple scale dynamics in proteins probed at multiple time scales through fluctuations of NMR chemical shifts," *The Journal of Physical Chemistry. B*, vol. 118, no. 14, pp. 3823–3831, 2014.

[49] G. L. Fan and Q. Z. Li, "Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 304, pp. 88–95, 2012.

[50] A. B. Sibley, M. Cosman, and V. V. Krishnan, "An empirical correlation between secondary structure content and averaged chemical shifts in proteins," *Biophysical Journal*, vol. 84, no. 2, pp. 1223–1227, 2003.

[51] R. T. Yang, C. Zhang, R. Gao, and L. N. Zhang, "A novel feature extraction method with feature selection to identify Golgi-resident protein types from imbalanced data," *International Journal of Molecular Sciences*, vol. 17, no. 2, p. 218, 2016.

[52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[53] L. Cheng, "Computational and biological methods for gene therapy," *Current Gene Therapy*, vol. 19, no. 4, pp. 210–210, 2019.

[54] L. Cheng and Y. Hu, "Human disease system biology," *Current Gene Therapy*, vol. 18, no. 5, pp. 255-256, 2018.

[55] W. X. Su, Q. Z. Li, L. Q. Zhang et al., "Gene expression classification using epigenetic features and DNA sequence composition in the human embryonic stem cell line H1," *Gene*, vol. 592, no. 1, pp. 227–234, 2016.

[56] B. Manavalan, T. H. Shin, and G. Lee, "PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine," *Frontiers in Microbiology*, vol. 9, p. 476, 2018.

[57] H. Y. Lai, Z. Y. Zhang, Z. D. Su et al., "iProEP: a computational predictor for predicting promoter," *Molecular Therapy-Nucleic Acids*, vol. 17, pp. 337–346, 2019.

[58] X. Li, Q. Tang, H. Tang, and W. Chen, "Identifying antioxidant proteins by combining multiple methods," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 858, 2020.

[59] W. Zhao, G. P. Li, J. Wang, Y. K. Zhou, Y. Gao, and P. F. Du, "Predicting protein sub-Golgi locations by combining functional domain enrichment scores with pseudo-amino acid compositions," *Journal of Theoretical Biology*, vol. 473, pp. 38–43, 2019.

[60] Y. H. Yang, C. Ma, J. S. Wang et al., "Prediction of N7-methylguanosine sites in human RNA based on optimal sequence features," *Genomics*, vol. 112, no. 6, pp. 4342–4347, 2020.

[61] M. L. Liu, W. Su, Z. X. Guan et al., "An overview on predicting protein subchloroplast localization by using machine learning methods," *Current Protein & Peptide Science*, vol. 21, 2020.

[62] Q. Tang, J. Kang, J. Yuan et al., "DNA4mC-LIP: a linear integration method to identify N4-methylcytosine site in multiple species," *Bioinformatics*, vol. 36, no. 11, pp. 3327–3335, 2020.

[63] J. Chen, J. Zhao, S. Yang, Z. Chen, and Z. Zhang, "Prediction of protein ubiquitination sites in Arabidopsis thaliana," *Current Bioinformatics*, vol. 14, no. 7, pp. 614–620, 2019.

[64] J.-H. Kuo, C.-C. Chang, C.-W. Chen, H.-H. Liang, C.-Y. Chang, and Y.-W. Chu, "Sequence-based structural B-cell epitope prediction by using two layer SVM model and association rule features," *Current Bioinformatics*, vol. 15, no. 3, pp. 246–252, 2020.

[65] Y. Jiao and P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications," *Quantitative Biology*, vol. 4, no. 4, pp. 320–330, 2016.

[66] F. M. Li and X. W. Gao, "Predicting gram-positive bacterial protein subcellular location by using combined features," *BioMed Research International*, vol. 2020, Article ID 9701734, 8 pages, 2020.

[67] L. Cheng, H. Zhuang, H. Ju et al., "Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a Mendelian randomization study," *Frontiers in Genetics*, vol. 10, 2019.

[68] L. Cheng, H. Zhao, P. Wang et al., "Computational methods for identifying similar diseases," *Molecular Therapy-Nucleic Acids*, vol. 18, pp. 590–604, 2019.

[69] L. Cheng, H. Zhuang, S. Yang, H. Jiang, S. Wang, and J. Zhang, "Exposing the causal effect of C-reactive protein on the risk of type 2 diabetes mellitus: a Mendelian randomization study," *Frontiers in Genetics*, vol. 9, p. 657, 2018.

[70] Z. Y. Zhang, Y. H. Yang, H. Ding, D. Wang, W. Chen, and H. Lin, "Design powerful predictor for mRNA subcellular location prediction in Homo sapiens," *Briefings in Bioinformatics*, 2020.

[71] F. Y. Dao, H. Lv, H. Zulfiqar et al., "A computational platform to identify origins of replication sites in eukaryotes," in *Briefings in Bioinformatics*, 2020.

[72] F. Y. Dao, H. Lv, Y. H. Yang, H. Zulfiqar, H. Gao, and H. Lin, "Computational identification of N6-methyladenosine sites in multiple tissues of mammals," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1084–1091, 2020.

[73] H. Yang, W. Yang, F. Y. Dao et al., "A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae," in *Briefings in Bioinformatics*, 2019.

[74] K. Liu and W. Chen, "iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications," *Bioinformatics*, vol. 36, no. 11, pp. 3336–3342, 2020.

[75] B.-Q. Li, Y.-H. Zhang, M.-L. Jin, T. Huang, and Y.-D. Cai, "Prediction of protein-peptide interactions with a nearest neighbor algorithm," *Current Bioinformatics*, vol. 13, no. 1, pp. 14–24, 2018.

[76] Z. Chen, P. Zhao, F. Li et al., "iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, 2018.

[77] R. Muhammod, S. Ahmed, D. M. Farid, S. Shatabda, A. Sharma, and A. Dehzangi, "PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences," *Bioinformatics*, vol. 35, no. 19, pp. 3831–3833, 2019.

[78] C. Ao, W. Zhou, L. Gao, B. Dong, and L. Yu, “Prediction of antioxidant proteins using hybrid feature representation method and random forest,” *Genomics*, vol. 112, no. 6, pp. 4666–4674, 2020.

[79] E. Kwon, M. Cho, H. Kim, and H. S. Son, “A study on host tropism determinants of influenza virus using machine learning,” *Current Bioinformatics*, vol. 15, no. 2, pp. 121–134, 2020.