Mathematical Biosciences and Engineering, 16(4): 2466–2480.
DOI: 10.3934/mbe.2019123
Received: 21 December 2018
Accepted: 04 March 2019
Published: 22 March 2019

http://www.aimspress.com/journal/MBE

Research article

Identification of hormone binding proteins based on machine learning methods

Jiu-Xin Tan[1], Shi-Hao Li[1], Zi-Mei Zhang[1], Cui-Xia Chen[2,3], Wei Chen [1,4,*], Hua Tang[5,*] and Hao Lin[1,*]

[1] Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
[2] National Research Institute for Family Planning, Beijing 100081, China
[3] National Center of Human Genetic Resources, Beijing 100081, China
[4] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China
[5] Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China

***Correspondence:** Email: chenweiimu@gmail.com; Tanghua771211@aliyun.com; hlin@uestc.edu.cn.

Abstract: The soluble carrier hormone binding protein (HBP) plays an important role in the growth of human and other animals. HBP can also selectively and non-covalently interact with hormone. Therefore, accurate identification of HBP is an important prerequisite for understanding its biological functions and molecular mechanisms. Since experimental methods are still labor intensive and cost ineffective to identify HBP, it's necessary to develop computational methods to accurately and efficiently identify HBP. In this paper, a machine learning-based method was proposed to identify HBP, in which the samples were encoded by using the optimal tripeptide composition obtained based on the binomial distribution method. In the 5-fold cross-validation test, the proposed method yielded an overall accuracy of 97.15%. For the convenience of scientific community, a user-friendly webserver called HBPred2.0 was built, which could be freely accessed at http://lin-group.cn/server/HBPred2.0/.

Keywords: hormone binding protein; tripeptide composition; binomial distribution method; feature selection; support vector machine; webserver

1. Introduction

Hormone-binding protein (HPB) is a kind of protein that selectively and non-covalently binds to hormone. HPB is a soluble outer region of the growth hormone receptor (HR), and is an important component of the growth hormone (GH)-insulin-like growth factor axis [1]. The abnormal expression of HBP can cause a variety of diseases [2]. Due to the complex in vivo effects of HBP, its biological function is still not fully understood [1]. Therefore, accurate identification of HBP will be helpful to understand the molecular mechanisms and regulatory pathways of HBP.

Traditional methods to identify HBP were wet biochemical experiments, such as immunoprecipitation, chromatography, crosslinking assays, etc [3–6]. However, the disadvantages of these methods, such as time-consuming and expensive, make them are unable to keep up with the rapid growth of protein sequences in the post-genomic era. Therefore, it is necessary to develop automatics machine learning methods to identify HBP. As a pioneer work, Tang et al. developed a support vector machine-based method to identify HBP in which proteins were encoded using the optimal features obtained by adopting optimized dipeptide composition [7]. Subsequently, Basith et al. developed a computational predictor named iGHBP, in which an optimal feature set was obtained based on combining dipeptide composition and amino acid index value by adopting two-step feature selection protocol [8]. However, the overall accuracy was still far from satisfactory. In order to improve the performance for the identification of HBP, it is necessary to apply new feature extraction and selection methods to select optimal features to represent HBP.

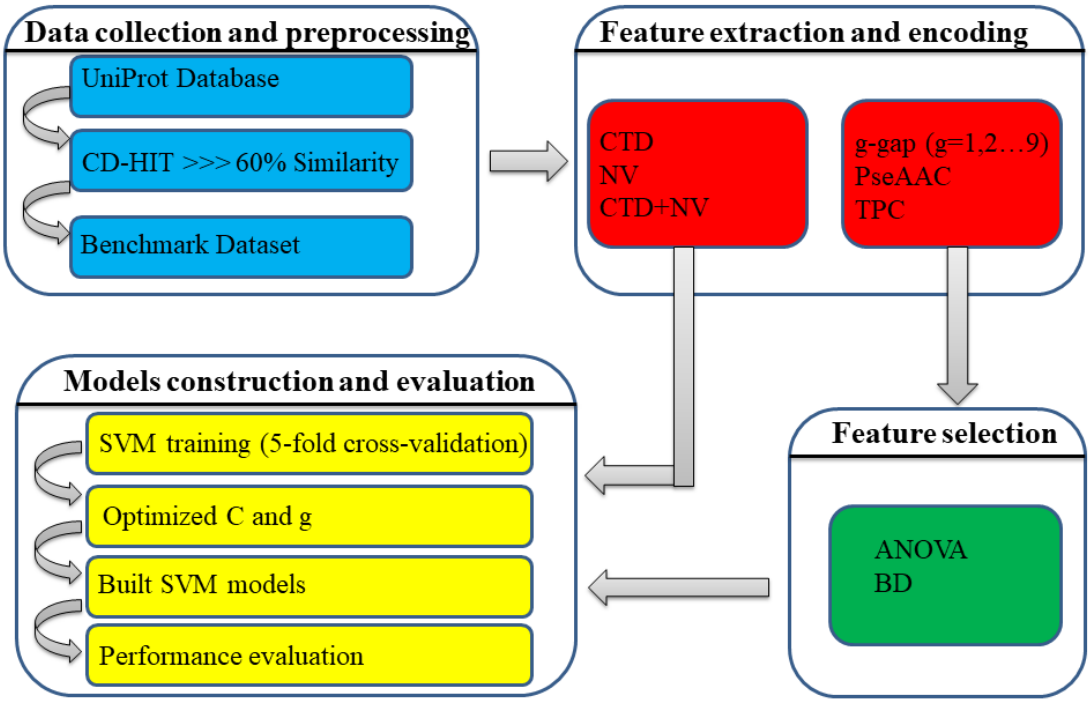

Figure 1. The framework of this work.

In this paper, by examining 5 feature encoding methods and 2 feature selection methods, we investigated the advantages and disadvantages of various models for identifying HBP and then

established a predictor called HBPred2.0 based on the optimal model. Finally, a user-friendly webserver was established for HBPred2.0. The paper is organized based on the following aspects (Figure 1): (1) The construction of benchmark dataset, (2) feature extraction and selection, (3) machine learning method, and (4) performance evaluation.

2. Materials and methods

2.1. Benchmark dataset and independent dataset

This paper adopted the benchmark dataset built by Tang et al. [7]. In the database, there are 123 hormone-binding proteins (HBPs) and 123 none hormone-binding proteins (non-HBPs). To verify the portability and validity of the model, we built a high quality independent dataset by obeying following rules. Firstly, we selected the 357 manually annotated and reviewed HBP proteins from Universal Protein Resource (UniProt) [9] using 'hormone-binding' as keywords in molecular function item of Gene Ontology. Subsequently, we excluded the proteins with sequence identity > 60% by using CD-HIT [10]. Thirdly, sequences that appear in the training dataset were excluded. As a result, 46 HBPs were obtained as independent positive samples. Negative samples were randomly selected from UniProt while using 'hormone' and 'DNA damage binding' as keywords in molecular function item of Gene Ontology, respectively. The sequence identities of negative samples are also $\leq$ 60%. Finally, 46 non-HBPs (37 hormone proteins and 9 DNA damage binding proteins) were randomly obtained. It should be noted that there is no similar sequences between the training and testing data. All data could be downloaded from http://lin-group.cn/server/HBPred2.0/download.html.

2.2. Feature extraction methods

2.2.1 Natural vector method (NV)

Suppose a sample protein **P** with *L* residues, it can be expressed as below.

$$\mathbf{P} = R_1 R_2 \ldots R_i \ldots R_L \tag{1}$$

where R_i represents the *i*-th amino acid residue of the sample protein **P**; $i = (1,2, \ldots L)$. The Natural Vector Method (NV) method is briefly described as follows [11]:

For each of the 20 amino acid *k*, define:

$$w_k(\cdot): (A, C, D, E, \ldots, W, Y) \rightarrow (0, 1) \tag{2}$$

where $w_k(R_i) = 1$, if $R_i = k$. otherwise, $w_k(R_i) = 0$.

Let n_k be the number of amino acid *k* in the protein sequence **P**, which can be calculated as:

$$n_k = \sum_{i=1}^{L} w_k(R_i) \tag{3}$$

Let $s_{(k)(i)}$ be the distance from the first amino acid (regarded as origin) to the *i*-th amino acid *k* in the protein sequence. Let T_k be the total distance of each set of the 20 amino acids. Let μ_k be the mean position of the amino acid *k*. And they can be calculated as:

$$\begin{cases} s_{(k)(i)} = i \times w_k(R_i) \\ T_k = \sum_{i=1}^{n_k} s_{(k)(i)} \\ \mu_k = T_k / n_k \end{cases} \tag{4}$$

Let D_2^k be the second-order normalized central moments, which can be calculated as:

$$D_2^k = \sum_{i=1}^{n_k} \frac{(s_{(k)(i)} - \mu_k)^2}{n_k \times L} \tag{5}$$

Thus, a sample protein **P** can be formulated as:

$$\mathbf{P} = [n_A, \mu_A, D_2^A, \ldots, n_R, \mu_R, D_2^{R_i}, \ldots n_Y, \mu_Y, D_2^Y]^T \tag{6}$$

where the symbol *T* is the transposition of the vector.

2.2.2. Composition transition distribution (CTD)

The CTD was first proposed for protein folding class prediction by Dubchak et al. in 1995 [12]. It's a global composition feature extraction method includes hydrophobicity, polarity, normalized van der Waals volume, polarizability, predicted secondary structure, solvent accessibility and so on. In this method, 20 amino acids were divided into 3 different groups: polar, neutral, and hydrophobic. For each of the amino acids attributes, three descriptors (C, T, D) were calculated. 'C' stands for 'Composition', which represents the composition percentage of each group in the peptide sequence, and thus can yield 3 features. 'T' stands for 'Transition', which represents the transition probability between two neighboring amino acids belonging to two different groups, and thus can yield 3 features. 'D' stands for 'Distribution', which represents the position (the first, 25%, 50%, 75%, or 100%) of amino acids in each group in the protein sequence, and thus can yield 5 features for each group (total 15 features).

In this paper, the sequence description of a sample protein **P** in term of hydrophobicity consists of 3 + 3 + 15 = 21 features.

2.2.3. G-gap dipeptide composition (*g*-gap)

Adjacent dipeptide composition can only express the correlation between two adjacent amino acid residues. In fact, the amino acids with *g*-gap residues may be adjacent in three-dimensional space [13]. To find important correlations in protein sequences, we used the *g*-gap dipeptide composition that extends from adjacent dipeptides. A protein **P** can be formulated as below by using this method.

$$\mathbf{P} = \left[v_1^g, v_2^g, \ldots, v_i^g, \ldots v_{400}^g\right]^T \tag{7}$$

where the symbol *T* is the transposition of the vector; the v_i^g is the frequency of the *i*-th (*i* = 1,2,…,400) *g*-gap dipeptide and can be formulated as:

$$v_i^g = \frac{n_i^g}{L-g-1} \tag{8}$$

where n_i^g is the number of the *i*-th *g*-gap dipeptide; *L* is the length of the protein **P**; *g* is the number of amino acid residues separated by two amino acid residues.

In this paper, we studied the cases of *g* ranging from 1 to 9 because the case of *g* = 0 has been studied in reference [7].

2.2.4. Pseudo amino acid composition (PseAAC)

The PseAAC method can not only include amino acid composition, but also the correlation of physicochemical properties between two residues [14,15]. In this paper, we adopted the type Ⅱ PseAAC, in which a sample protein **P** can be formulated as below.

$$\mathbf{P} = [x_1, x_2 \ldots, x_{400}, x_{401}, \ldots x_{400+9\lambda}]^T \tag{9}$$

where '9' is the number of amino acid physicochemical properties considered, namely, hydrophobicity, hydrophilicity, mass, pK1, pK2, pI, rigidity, flexibility and irreplaceability; 'λ' is the rank of correlation; 'x' is the frequencies for each element and is formulated as:

$$x_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{400} f_u + \omega \sum_{j=1}^{9\lambda} \tau_j}, (1 \le u \le 400) \\ \frac{\omega \tau_j}{\sum_{i=1}^{400} f_u + \omega \sum_{j=1}^{9\lambda} \tau_j}, (401 \le u \le 400 + 9\lambda) \end{cases} \tag{10}$$

where ω is the weight factor for the sequence order effect; f_u is the frequency of the 400 dipeptides; τ_j is the correlation factor of the physicochemical properties between residues. More detailed information about the formula derivation process can be found in the reference [16].

In this paper, the parameter λ is from 1 to 95 with the step of 1, the parameter ω is from 0.1 to 1 with the step of 0.1. Therefore, 95×10 = 950 feature subsets based on PseAAC will be obtained.

2.2.5. Tripeptide composition (TPC)

Tripeptide is composed of three adjacent amino acids in a protein sequence, which is a biosignaling with minimal functionality. By adopting TPC, a sample protein **P** can be formulated by：

$$\mathbf{P} = [t_1, t_2, \ldots t_i, \ldots, t_{8000}]^T \tag{11}$$

where the symbol *T* is the transposition of the vector; the t_i is the frequency of the *i*-th (*i* = 1,2,…,8000) tripeptide and can be formulated as:

$$t_i = \frac{n_i}{L-2} \tag{12}$$

where n_i is the number of the *i*-th tripeptide; *L* is the length of the protein **P**.

2.3. Feature selection methods

2.3.1. Analysis of variance (ANOVA)

Feature selection is important to improve the classification performance. It can filter the noisy features [17–20]. We adopted the ANOVA method to select optimal features from *g*-gap dipeptide compositions and PseAAC. The ANOVA method calculated the ratio of the variance among groups and the variance within groups for each attribute [21,22]. The formula expressions can be described as follows:

$$F(i) = \frac{S_b^2(i)}{S_w^2(i)} \tag{13}$$

where $F(i)$ is the score of the *i*-th feature, a high $F(i)$-value means a high ability to identify the sample; $S_w^2(i)$ is the variance within groups; $S_b^2(i)$ is the variance among groups; and they can be calculated as follows:

$$\begin{cases} S_b^2(i) = \frac{SS_b(i)}{K-1} \\ S_w^2(i) = \frac{SS_w(i)}{N-K} \end{cases} \tag{14}$$

where $SS_b(i)$ is the sum of the squares between the groups; $SS_w(i)$ is the sum of squares within the groups; *K* is the total number of classes; *N* is the total number of samples.

2.3.2. Binomial distribution (BD)

We adopted the BD method to select optimal features from tripeptide composition [21]. In this algorithm, the confidence level (*CL*) of each feature can be calculated by:

$$CL_{ij} = 1 - \sum_{k=n_{ij}}^{N_i} \frac{N_i!}{k!(N_i-k)!} q_j^k (1-q_j)^{N_i-k} \tag{15}$$

where CL_{ij} is the confidence level for the *i*-th tripeptide in the *j*-th type; *j* denotes the type of samples (positive sample or negative sample); N_i is the total number of the *i*-th tripeptide in the dataset; the probability q_j is the relative frequency of type *j* in the dataset;

According to the formula as defined in Eq. (15), a high *CL*-value means a high ability to identify the sample. The BD method can extract the over-represented motifs, which is an excellent statistical method widely used in bioinformatics [23,24].

2.3.3. Incremental feature selection (IFS) process

In general, if a model was built on a low-dimensional feature subset, it will not provide enough information. On the contrary, if a model was built on a high-dimensional feature subset, it can lead to information redundancy and overfitting problems. Therefore, the ANOVA and BD method with the IFS process and 5-fold cross-validation was applied to investigate the optimal feature set with the maximum accuracy [7,25–27] (Figure 2). We ranked all features according to the *F*(*i*)-values or *CL*-values and obtained new feature vectors, which are shown below.

$$\mathbf{P}' = [g'_1, g'_2, \ldots g'_n]^T \tag{16}$$

The first feature subset contains the feature with the highest *F(i)*-value or *CL*-value, $\mathbf{P}' = [g'_1]^T$; By adding the second highest *F(i)*-value or *CL*-value to the first subset, the second feature subset $\mathbf{P}' = [g'_1, g'_2]^T$ is formed. The procedure was repeated until all features were considered.

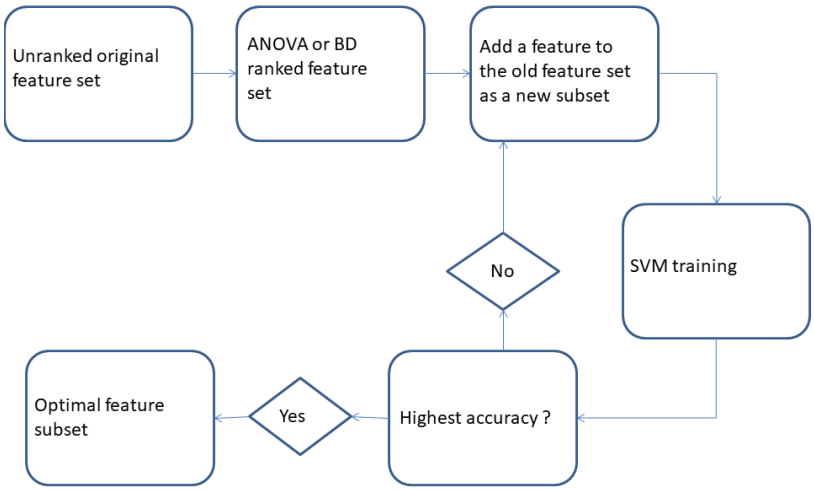

Figure 2. The framework of the IFS process.

2.4. Support vector machine (SVM)

The support vector machine (SVM) is a supervised machine learning method and has been widely used in bioinformatics [28–33]. Its main idea is to map the input features from low-dimensional space to a high-dimensional space through nonlinear transformation and find the optimal linear classification surface. For convenience, SVM software packages LibSVM can be download from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. In the current study, the LibSVM-3.22 package was adopted to investigate the performance for identifying HBP. Besides, the radical basis function kernel was selected to perform predictions. The grid search spaces are [2^{-5}, 2^{15}] with step of 2 for penalty parameter *C* and [2^{3}, 2^{-15}] with step of 2^{-1} for kernel parameter *g*.

2.5. Performance evaluation

Three cross-validation methods, namely, the independent dataset test, the sub-sampling test, and the jackknife test, are widely used to investigate the performance of a predictor in practical application [30,34–41]. In order to save computing time, the 5-fold cross-validation test was adopted to calculate the optimal parameter C and g of SVM in this paper.

Five evaluation indexes were adopted to evaluate the models [42–49]. Sensitivity (S_n) is used to evaluate the model's ability to correctly predict positive samples. Specificity (S_p) is used to evaluate the model's ability to correctly predict negative samples. Overall Accuracy (Acc) reflects the proportion of the entire benchmark dataset that can be correctly predicted. The Matthew correlation coefficient (Mcc) is used to evaluate the reliability of the algorithm. Area under the ROC curve (AUC) reflects model's classification ability across decision values. They can be calculated as follows:

$$\begin{cases} S_n = \frac{TP}{TP+FN} \\ S_p = \frac{TN}{TN+FP} \\ Acc = \frac{TP+TN}{TP+TN+FN+FP} \\ Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \tag{16}$$

where *TP*, *TN*, *FP*, and *FN* represent the number of the correctly recognized positive samples, the number of the correctly recognized negative samples, the number of negative samples recognized as positive samples, and the number of positive samples recognized as negative samples, respectively.

3. Results and discussion

3.1. Performances of different features

In this study, we examined the performance of 5 feature extraction methods and their combinations. Based on CTD, NV, CTD+NV methods, protein samples can be expressed as 21-D (dimensional), 60-D and 81-D vector, respectively. The Accs of 60.16%, 70.33% and 67.07% were obtained by using SVM in the 5-fold cross-validation, respectively (as shown in Table 1). It was found that the prediction performances were far from satisfactory.

Based on the *g*-gap method, a protein sample can be expressed as a 400-D vector. By changing the value of *g* from 1 to 9, we obtained 9 feature subsets. Firstly, we investigated the performances of these 400-D features subsets based on SVM. The results were reported in Figure 3A. Subsequently, the ANOVA method with the IFS process was applied to investigate the optimal feature set, and the results were recorded in Figure 3B. One may notice that while g = 1, a maximum Acc of 80.89% was obtained when the top 144 features were used. Obviously, Accs were significantly increased by adopting ANOVA method. However, prediction performances still needed to improve.

Based on the PseAAC method, we obtained 95×10 = 950 (95 kinds of λ and 10 kinds of ω) feature subsets. Firstly, we investigated the performances of these 950 models by using SVM in the 5-fold cross-validation test and reported the results in Figure 4A. It was found that the maximum Acc of 76.83% was achieved when λ = 18 and ω = 0.1. In order to improve Acc, the ANOVA method was adopted to rank the $400 + 18 \times 9 = 572$ features. By adopting SVM with IFS, a maximum Acc of 84.15% was obtained when the top 194 features were used (Figure 4B). Although the result was encouraging, the Acc still has room to rise.

Table 1. The results and the corresponding number of features based on different methods.

Feature extraction	C	g	S_n(%)	S_p(%)	Acc(%)	Mcc	AUC
CTD (21-D)	2	2^3	36.59	83.74	60.16	0.230	0.654
NV (60-D)	2^{-5}	2^{-13}	70.73	69.92	70.33	0.407	0.762
CTD+NV (81-D)	2^9	2^{-7}	70.73	63.41	67.07	0.342	0.709

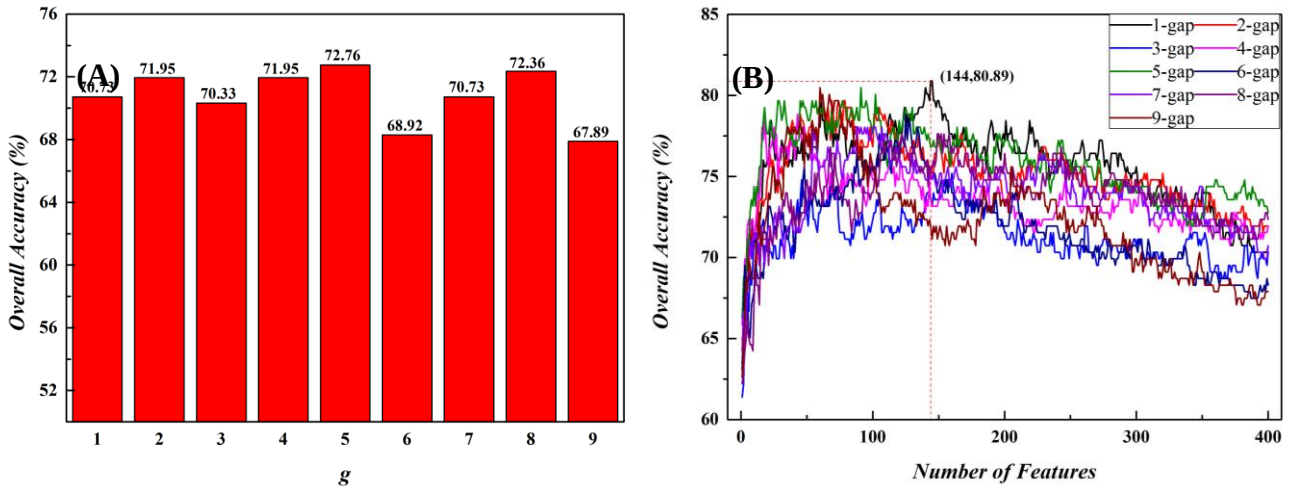

Figure 3. *Acc*s for g-gap dipeptide composition. **(A)** Different g values corresponding to different *Acc*s; **(B)** A plot showing the IFS curves based on g-gap methods.

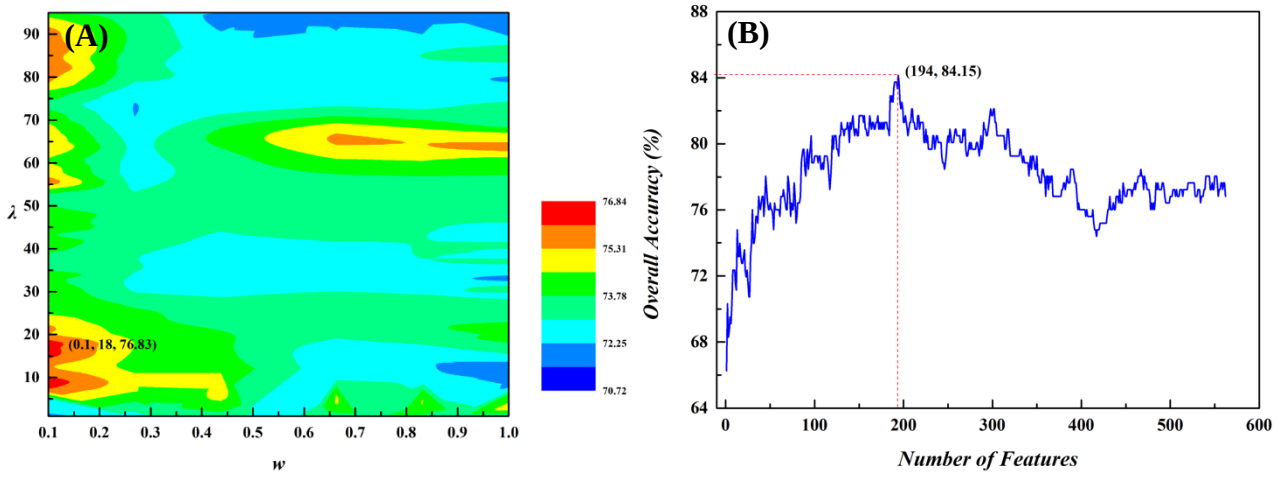

Figure 4. *Acc*s for g-gap dipeptide composition. **(A)** A heat map for the *Acc*s of 950 PseAAC models; **(B)** A plot showing the IFS curve based on PseAAC method.

Based on the TPC method, 8000 features were extracted for each protein sequence. Considering that it would lead to overfitting problem, the BD method was adopted as the feature selection method. By adopting SVM with IFS process in the 5-fold cross-validation test, a maximum *Acc* of 97.15% was obtained when the top 1169 features were used (Figure 5). In this case, the S_n, S_p and *Mcc* are 96.75%, 97.56%, and 0.943, respectively The AUC reached 0.994, this result indicates that the performance of the model based on the optimal TPC is smart and reliable for identifying HBP.

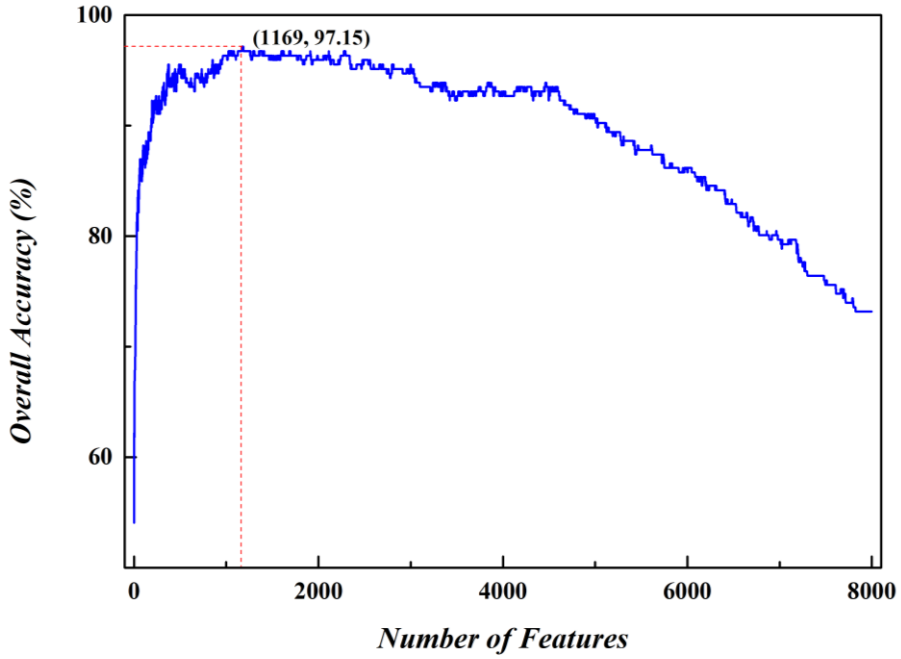

Figure 5. A plot showing the IFS curve based on TPC method.

3.2. Comparison with other methods

In order to show the superiority of SVM to identify HBP, we compared its performance with those of other machine learning algorithms based on the same feature subset (i.e. 1169 optimal features). From Table 2, we can find that the SVM classifier could produce the best performance among these algorithms. Thus, the final model was constructed based on SVM.

Table 2. Comparing SVM with other classifiers.

Classifier	S_n (%)	S_p (%)	Acc (%)	Mcc	AUC
J48	63.41	56.91	60.16	0.204	0.601
Bagging	80.49	57.72	69.11	0.392	0.770
Random Forest	88.62	84.55	86.59	0.732	0.945
Naive Bayes	95.93	92.68	94.31	0.887	0.965
SVM	96.75	97.56	97.15	0.943	0.994

It is also necessary to compare the methods proposed in this paper with existing methods. Table 3 shows the detailed results of different methods for identifying HBP. Based on the same benchmark dataset, Tang et al. achieved an Acc of 84.9% by using a SVM-based method, in which proteins sequences were encoded using the optimal 0-gap dipeptide composition features obtained by the ANOVA feature selection technique [7]. Basith et al. obtained an Acc of 84.96% in cross-validation test by training an extremely randomized tree with optimal features obtained from dipeptide composition and amino acid index values based on two-step feature selection [8]. Our proposed method could produce an Acc of 97.15% which is superior to the two published results, demonstrating that our method is more powerful for identifying HBP.

Table 3. Comparing our method with other published methods.

Reference	Methods	S_n (%)	S_p (%)	Acc (%)	Mcc	AUC
[7]	HBPred	88.6	81.3	84.9	-	-
[8]	iGHBP	88.62	81.30	84.96	-	0.701
This work	HBPred2.0	96.75	97.56	97.15	0.943	0.994

3.3. Performance evaluation based on the independent dataset

For further comparing the performance of these methods, an independent dataset was used. The results were recorded in Table 4. One may observe that the HBPred2.0 predictor achieved the best performance among the three predictors, suggesting that HBPre2.0 has better generalization ability.

Table 4. Performance evaluation based on the independent dataset.

Reference	Methods	S_n (%)	S_p (%)	Acc (%)	Mcc	AUC
[7]	HBPred	80.43	56.52	68.48	0.381	0.714
[8]	iGHBP	86.96	47.83	67.39	0.380	-
This work	HBPred2.0	89.13	80.43	84.78	0.698	0.814

Specificity could reflect the discriminated capability of model on negative samples. From the Table 4, a higher specificity of the HBPred2.0 indicates that the model could produce less false positives.

4. Conclusion

In this paper, we systematically investigated the performances of various features and classifiers on HBP prediction. By a great number of experiments, we obtained the best model by combining SVM with optimal tripeptide composition. This model could produce the overall accuracy of 84.78% on the independent data. Finally, Due to published database [50–53] and webserver [54–63] could provide more convenience for scientific community, we established a free webserver for the proposed method, called HBPred2.0, which can be free accessed form http://lin-group.cn/server/HBPred2.0/. We expect that the tool will help scholars to study the mechanism of HBP's function, and promote the development of related drug research.

Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (61772119, 31771471, 61702430), Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), the Central Public Interest Scientific Institution Basal Research Fund (No. 2018GJM06).

Conflicts of interest

The authors declare that there is no conflict of interest.

References

1. G. Baumann, Growth hormone binding protein. The soluble growth hormone receptor, *Minerva. Endocrinol.*, **27** (2002), 265–276.
2. J. A. Kraut and N. E. Madias, Adverse effects of the metabolic acidosis of chronic kidney disease, *Adv. Chronic. Kidney Dis.*, **24** (2017), 289–297.
3. F. Sohm, I. Manfroid and A. Pezet, et al., Identification and modulation of a growth hormone-binding protein in rainbow trout (Oncorhynchus mykiss) plasma during seawater adaptation, *Gen. Comp. Endocrinol.*, **111** (1998), 216–224.
4. Y. Zhang and T. A. Marchant, Identification of serum GH-binding proteins in the goldfish (Carassius auratus) and comparison with mammalian GH-binding proteins, *J. Endocrinol.*, **161** (1999), 255–262.
5. I. E. Einarsdottir, N. Gong and E. Jonsson, et al., Plasma growth hormone-binding protein levels in Atlantic salmon Salmo salar during smoltification and seawater transfer, *J. Fish Biol.*, **85** (2014), 1279–1296.
6. S. Fisker, J. Frystyk and L. Skriver, et al., A simple, rapid immunometric assay for determination of functional and growth hormone-occupied growth hormone-binding protein in human serum, *Eur. J. Clin. Invest.*, **26** (1996), 779–785.
7. H. Tang, Y. W. Zhao and P. Zou, et al., HBPred: a tool to identify growth hormone-binding proteins, *Int. J. Biol. Sci.*, **14** (2018), 957–964.
8. S. Basith, B. Manavalan and T. H. Shin, et al., iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree, *Comput. Struct. Biotechnol. J.*, **16** (2018), 412–420.
9. L. Breuza, S. Poux and A. Estreicher, et al., The UniProtKB guide to the human proteome, *Database (Oxford)*, **2016** (2016).
10. L. Fu, B. Niu and Z. Zhu, et al., CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, **28** (2012), 3150–3152.
11. K. Tian, X. Zhao and S. S. Yau, Convex hull analysis of evolutionary and phylogenetic relationships between biological groups, *J.Theor. Biol.*, **456** (2018), 34–40.
12. I. Dubchak, I. Muchnik and S. R. Holbrook, et al., Prediction of protein folding class using global description of amino acid sequence, *Proc. Natl. Acad. Sci. U S A*, **92** (1995), 8700–8704.
13. H. Tang, W. Chen and H. Lin, Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique, *Mol. Biosyst.*, **12** (2016), 1269–1275.
14. K. C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, *J. Theor. Biol.*, **273** (2011), 236–247.
15. K. C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, *Proteins*, **43** (2001), 246–255.
16. F. Y. Dao, H. Yang and Z. D. Su, et al., Recent advances in conotoxin classification by using machine learning methods, *Molecules*, **22** (2017), in press.
17. Q. Zou, S. Wan and Y. Ju, et al., Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy, *BMC System. Biol.*, **10** (2016), 114.
18. L. Wei, R. Su and B. Wang, et al., Integration of deep feature representations and handcrafted features to improve the prediction of N^6-methyladenosine sites, *Neurocomputing*, **324** (2019), 3–9.

19. G. H. Huang and J. C. Li, Feature extractions for computationally predicting protein post-translational modifications, *Curr. Bioinform.,* **13** (2018), 387–395.
20. Q. Zou, J. Zeng and L. Cao, et al., A novel features ranking metric with application to scalable visual and bioinformatics data classification, *Neurocomputing*, **173** (2016), 346–354.
21. H. Y. Lai, X. X. Chen and W. Chen, et al., Sequence-based predictive modeling to identify cancerlectins, *Oncotarget*, **8** (2017), 28169–28175.
22. X. X. Chen, H. Tang and W. C. Li, et al., Identification of bacterial cell wall lyases via pseudo amino acid composition, *Biomed. Res. Int.*, **2016** (2016), 1654623.
23. X. J. Zhu, C. Q. Feng and H. Y. Lai, et al., Predicting protein structural classes for low-similarity sequences by evaluating different features, *Knowled. System.*, **163** (2019), 787–793.
24. H. Yang, W. R. Qiu and G. Q. Liu, et al., iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC, *Int. J. Biol. Sci.*, **14** (2018), 883–891.
25. H. Yang, H. Tang and X. X. Chen, et al., Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition, *Biomed. Res. Int.*, **2016** (2016), 5413903.
26. C. Q. Feng, Z. Y. Zhang and X. J. Zhu, et al., iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators, *Bioinformatics*, (2018), in press.
27. F. Y. Dao, H. Lv and F. Wang, et al., Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique, *Bioinformatics*, (2018), in press.
28. H. Lin, Z. Y. Liang and H. Tang, et al., Identifying sigma70 promoters with novel pseudo nucleotide composition, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, (2017), in press.
29. W. Chen, H. Yang and P. Feng, et al., iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, *Bioinformatics*, **33** (2017), 3518–3523.
30. W. Chen, P. Feng and T. Liu, et al., Recent advances in machine learning methods for predicting heat shock proteins, *Curr. Drug Metab.*, (2018), in press.
31. D. Li, Y. Ju and Q. Zou, Protein folds prediction with hierarchical structured SVM, *Curr. Proteom.*, **13** (2016), 79–85.
32. N. Zhang, S. Yu and Y. Guo, et al., Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine, *Curr. Bioinform.*, **13** (2018), 50–56.
33. H. Yang, H. Lv and H. Ding, et al., iRNA-2OM: A sequence-based predictor for identifying 2'-o-methylation sites in homo sapiens, *J. Comput. Biol.*, **25** (2018), 1266–1277.
34. P. M. Feng, H. Ding and W. Chen, et al., Naive Bayes classifier with feature selection to identify phage virion proteins, *Comput. Math. Methods Med.*, **2013** (2013), 530696.
35. B. Manavalan, S. Subramaniyam and T. H. Shin, et al., Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy, *J. Proteom. Res.*, **17** (2018), 2715–2726.
36. P. M. Feng, W. Chen and H. Lin, et al., iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, *Anal. Biochem.*, **442** (2013), 118–125.
37. P. M. Feng, H. Lin and W. Chen, Identification of antioxidants from sequence information using naive Bayes, *Comput. Math. Method. Med.*, **2013** (2013), 567529.
38. P. Feng, H. Yang and H. Ding, et al., iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC,

Genomics, (2018), in press.
39. W. Chen, P. M. Feng and E. Z. Deng, et al., iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, *Anal. Biochem.*, **462** (2014), 76–83.
40. L. Z. Yuan, E. F. Yong and Z. Wei, et al., Using quadratic discriminant analysis to predict protein secondary structure based on chemical shifts, *Curr. Bioinform.*, **12** (2017), 52–56.
41. W. Chen, H. Lv, and F. Nie, et al., i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome, *Bioinformatics*, (2019), in press.
42. Y. Bao, S. Marini and T. Tamura, et al., Toward more accurate prediction of caspase cleavage sites: a comprehensive review of current methods, tools and features, *Brief Bioinform.*, (2018), in press.
43. H. Tang, C. M. Zhang and R. Chen, et al., Identification of secretory proteins of malaria parasite by feature selection technique, *Letter. Organic Chem.*, **14** (2017), 621–624.
44. H. Tang, R. Z. Cao and W. Wang, et al., A two-step discriminated method to identify thermophilic proteins, *Int. J. Biomath.*, **10** (2017), in press.
45. S. Patel, R. Tripathi and V. Kumari, et al., DeepInteract: Deep neural network based protein-protein interaction prediction tool, *Curr. Bioinform.*, **12** (2017), 551–557.
46. R. Z. Cao, B. Adhikari and D. Bhattacharya, et al., QAcon: single model quality assessment using protein structural and contact information with machine learning techniques, *Bioinform.*, **33** (2017), 586–588.
47. R. Cao, C. Freitas and L. Chan, et al., ProLanGO: Protein function prediction using neural machine translation based on a recurrent neural network, *Molecules*, **22** (2017), in press.
48. B. Manavalan, T. H. Shin and M. O. Kim, et al., PIP-EL: A new ensemble learning method for improved proinflammatory peptide predictions, *Front. Immunol.*, **9** (2018), 1783.
49. B. Manavalan, T. H. Shin and G. Lee, PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine, *Front. Microbiol.*, **9** (2018), 476.
50. T. Cui, L. Zhang and Y. Huang, et al., MNDR v2.0: an updated resource of ncRNA-disease associations in mammals, *Nucleic Acids Res.*, **46** (2018), D371–D374.
51. T. Zhang, P. Tan and L. Wang, et al., RNALocate: a resource for RNA subcellular localizations, *Nucleic Acids Res.*, **45** (2017), D135–D138.
52. Y. Yi, Y. Zhao and C. Li, et al., RAID v2.0: an updated resource of RNA-associated interactions across organisms, *Nucleic Acids Res.*, **45** (2017), D115–D118.
53. Z.Y. Liang, H.Y. Lai and H. Yang, et al., Pro54DB: a database for experimentally verified sigma-54 promoters, *Bioinformatics*, **33** (2017), 467–469.
54. J. Song, Y. Wang and F. Li, et al., iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, *Brief Bioinform.*, (2018), in press.
55. J. Song, F. Li and A. Leier, et al., PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, *Bioinformatics*, **34** (2018), 684–687.
56. R. Cao and J. Cheng, Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks, *Methods*, **93** (2016), 84–91.
57. W. Chen, P.M. Feng and E.Z. Deng, et al., iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, *Anal. Biochem.*, **462** (2014), 76–83.
58. I. Naseem, S. Khan and R. Togneri, et al., ECMSRC: A sparse learning approach for the

prediction of extracellular matrix proteins, *Curr. Bioinform.*, **12** (2017), 361–368.
59. R. Z. Cao, D. Bhattacharya and J. Hou, et al., DeepQA: improving the estimation of single protein model quality with deep belief networks, *BMC Bioinform.*, **17** (2016), in press.
60. B. Manavalan, S. Basith and T. H. Shin, et al., MLACP: machine-learning-based prediction of anticancer peptides, *Oncotarget*, **8** (2017), 77121–77136.
61. B. Manavalan, S. Basith and T. H. Shin, et al., mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, *Bioinformatics*, (2018), in press.
62. B. Manavalan, R. G. Govindaraj and T. H. Shin, et al., iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction, *Front. Immunol.*, **9** (2018), 1695.
63. B. Manavalan, T. H. Shin and G. Lee, DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest, *Oncotarget*, **9** (2018), 1944–1956.

AIMS Press