VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2021, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 9, Number 2, April-June 2021 *pp: 30-43*

Identification of Hormone Binding Proteins via PseAAC

Muhammad Moshin Ali[1], Muhammad Adeel Ashraf[1], SALMAN QADRI

[1] Department of Computer Sciences, University of Management and Technology, Lahore, Pakistan
[2] Department of Computer Sciences, MNS University of Agriculture, Multan, Pakistan
*Corresponding Author: F2019108015@umt.edu.pk

ABSTRACT
Hormone binding protein (HBP) is analogous to a soluble protein carrier and can interact with human hormones non-covalently and selectively. HBP plays an imperative function in the growth of life, but its role remains uncertain. The first step in advancing the study of their work and recognizing their biological process is the correct identification of HBPs. It is difficult, however, to correctly classify HBPs from via conventional biochemical experiments, due to high experimental costs and long experimental time, more and more proteins. Meanwhile, experimental methods are still labor-intensive and cost-effective to identify HBP, developing computational methods to identify HBP accurately and efficiently is crucial. In this analysis, a method based on machine learning was suggested to classify the HBP during which the samples were encoded using the optimal composition of tripeptides obtained by supporting the binomial distribution method. The suggested approach yielded an overall precision of 97.15 percent in the 5-fold cross-validation test. A new technique for recognizing HBPs is provided by this report.

KEYWORDS
Identifying hormone-binding proteins; tripeptide composition (TPC); deep neural network; binomial distribution (BD) technique; machine learning (ML) method; feature selection

JOURNAL INFO
HISTORY: Received: April 20, 2021 Accepted: June 15, 2021 Published:.June 30, 2021.

INTRODUCTION

Hormone-binding protein (HPB) is like a protein that binds to hormone non-covalently and selectively and carries hormone to target tissue to achieve a desired effect (H. Tang Y. W., 2018). HBPs were first described in pregnant rabbit, human and mouse plasma a while ago (Shahid Akbar, 2020). They are concerned with controlling the supply of hormones in the circulatory system and have an effect on the metabolism or actions of other cells that have functional hormone receptors.

HBP are basically associated with targeted hormones which are then passed to specific cells to perform specific task and to make the required impact (G. Bauman, 2002,).

HBPs are attached with hormone supply in the human blood system that affects the performance of other tissues connected with hormones (G. Ozzola, 2016,).

sex HBPs produced primarily in the liver bind to sex steroid hormones and thus control their bioavailability (H. Tang Y. W., 2018). The abnormal expression of HBPs induces different diseases at all times (Madias, 2017) Testosterone, acromegaly, Graves' disease. Therefore, clarifying the role of HBPs and their mechanisms of regulation is important.

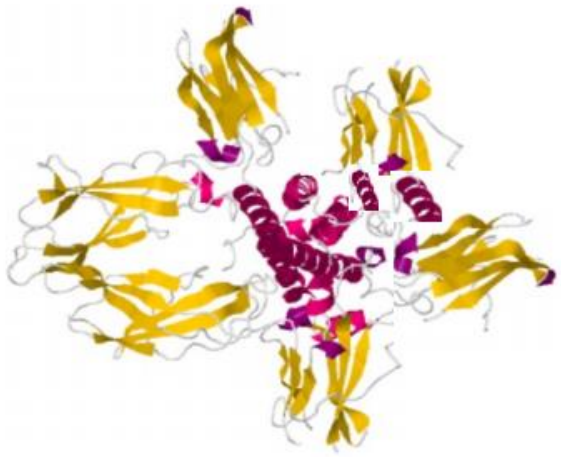

Figure 1. Schematic diagram of human hormone (red) to two HBPs (yellow)

HPB can be a soluble growth hormone receptor (HR) outer region and is an essential component of the axis of growth hormone (GH)-insulin-like growth factor (Jiu-Xin Tan1, 2019). Its biological role is still not well understood because of the diverse in vivo effects of HBP (Shahid Akbar, 2020).

Precise identification of HBP would also be useful in understanding HBP's molecular mechanisms and regulatory pathways

Several standard experimental approaches have recently been used to classify HBPs (Y. Zhang, 1999). Traditional

biochemical methods, however, are time consuming and costly, which makes them ineffective because of the rapid increase in the area's protein sequences. That is why, to recognize HBPs, an intelligent and automated machine learning model is highly needed (Shahid Akbar, 2020). Wet biochemical studies, such as immunoprecipitation, chromatography, crosslinking assays, etc., have been conventional methods to identify HBP (I.E. Einarsdottir, 2014).

However, in the post-genomic period, the drawbacks of these techniques, such as time-consuming and costly, render them unable to keep up with the rapid growth of protein sequences (Jiu-Xin Tan1, 2019). Consequently, the development of automated machine learning methods to classify HBPP is mandatory (Shahid Akbar, 2020). As the work of a visionary, Tang et al. developed a machine-based support vector system to classify HBP in which proteins were encoded by implementing optimized dipeptide composition using the optimal features obtained (H. Tang Y. W., 2018).

Another tool used by PsePSSM-based evolutionary features and deep learning approach to classify hormone binding proteins was used by (Shahid Akbar, 2020). Basith et al. successively developed a statistical predictor called iGHBP in which an optimal feature set was obtained by implementing a two-step feature selection protocol based on combining dipeptide composition and amino acid index value (S. Basith, 2018).

Wang et al used an ensemble classification learner technique to classify HBPs (Shahid Akbar, 2020). However, tripeptide composition (TPC) was used to obtain data from protein HBP sequences. But the accuracy overall was still far from fair.

It is important to apply new feature extraction and selection methods to select optimal features to reflect HBP in order to improve performance for the identification of HBP

In this analysis, we selected the methodology in which we investigated the pros and cons of different models to identifying HBP by examining 10 feature extraction (encoding) methods and 4 feature selection methods (Jiu-Xin Tan1, 2019), and then developed a predictor named HBPred2.0 based on the optimal model, which could be freely accessed at http://lin-group.cn/server/HBPred2.0/.

RELATED WORK

BENCHMARK DATASET

This study introduced the benchmark dataset developed by (H. Tang Y. W., 2018). 123 hormone-binding proteins (HBPs) and 123 hormone-binding proteins (non-HBPs) are included in the database (Jiu-Xin Tan1, 2019). We developed a high-quality independent dataset to check the portability and validity of the model by obeying the following laws. The 357 manually annotated and checked HBP proteins from Universal Protein Resource (UniProt) (L. Breuza, 2016) were initially nominated using 'hormone-binding' as keywords in the Gene Ontology molecular function item. Afterwards, by using CD-HIT (L. Fu, 2012) we removed the proteins with sequence identity > 60% [35-58]. Thirdly, there have been omitted sequences that occur in the training dataset. As a consequence, 46 HBPs were obtained as positive independent samples. After that, negative samples were randomly selected from UniProt while using 'hormone' and 'DNA damage binding' as keywords, respectively, in the Gene Ontology molecular function item. There are also 60% of the sequence identities of negative samples. In conclusion, 46 non-HBPs were acquired at random (37 hormone proteins and 9 DNA damage binding proteins). It should be noted that no identical sequences exist between the data from training and research.

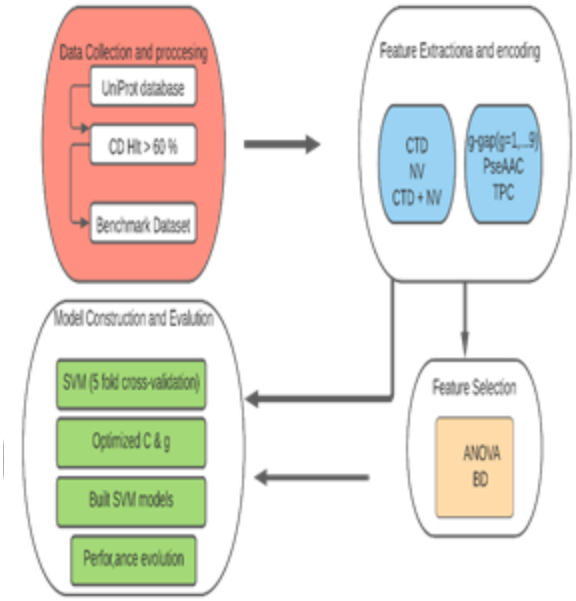

Figure 2. The background of work

FEATURE,EXTRACTION METHODS

NATURAL VECTOR METHOD (NV)

If the sample protein P with L residues is presumed, it can be expressed as below.

$$P = R1R2\ldots Ri\ldots RL \quad (1)$$

Where Ri reflects the sample protein P's i-th amino acid residue; I = (1,2,....L). The method of the natural vector method (NV) is defined briefly as follows (Jiu-Xin Tan1, 2019):

Define: for each of the 20 k amino acids:

$$wk(.) : (A,C,D,E,..,W,Y) \rightarrow (0,1) \quad (2)$$

Where $w_k(R_i)$ = 1, if R_i = k. almost, $w_k(R_i)$ = 0.

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

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

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 (Jiu-Xin Tan1, 2019). Let T_k be the whole detachment of each set of the 20 amino acids.
Let μ_k be the nasty location of the amino acid k. And they can be deliberate as:

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

Let D_2^k be the second-order normalized vital moments, which could be intended as:

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

Hence, a sample protein P can be uttered as:

$$\mathbf{P} = [n_A, \mu_A, D_2^A, \dots, n_R, \mu_R, D_2^{R_i}, \dots n_Y, \mu_Y, D_2^Y]^T \quad (6)$$

In the above equation symbol of T is the transposition of the vector.

Random Forest (RF*)*

RF is one of the most common ML processes, and utilizes hundreds or thousands of liberated verdict trees to perform arrangement and reversion (L., 2001,). RF incorporates the ideologies of bagging and random (S. Basith, 2018) function collection. For a given training data set (D), create a new training data set (D_i) by uniformly drawing N boot-strapped illustrations from D. (S. Basith, 2018)Grow a tree consuming D_i besides recurrence the following steps at each node of the tree until its fully grown: (i) pick mtry random-features after the original features and choice the finest flexible by optimizing the impurity criteria, and (ii) fragmented the node hooked on two child nodes (S. Basith, 2018). The hierarchy grows unless the amount of data in the node is below the specified threshold (nsplit). Repeat the above-mentioned steps to build a large quantity (ntree) of the saplings of organization. Input structures (S. Basith, 2018) are recycled to classify the test data. Passed from the root to each tree's end node based on Splits are predetermined. The concluding organization is considered as the plurality of class from the forest.

COMPOSITION TRANSITION DISTRIBUTION (CTD)

For protein folding class prediction, the CTD was first proposed (S. Basith, 2018). Hydrophobicity, polarity, normalized van der Waals length, polarizability, projected secondary structure, solvent accessibility, etc. (Jiu-Xin Tan1, 2019) are a global composition function extraction process. In our previous study, a detailed overview of computational CTD features was given. 20 amino acids were divided into 3 different classes in this method: polar, neutral, and hydrophobic. There were three descriptors (C, T, D) determined for each of the amino acid attributes. 'C' stands for 'Composition' that signifies the composition percentage of separately group in the peptide sequence, and thus can yield 3 features. 'T' stands for 'Transition', which signifies 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%) (Jiu-Xin Tan1, 2019) of amino acids in each group in the protein sequence, and thus can yield 5 features for each group (total 15 features). In this research of identify, the sequence description of a sample protein P in term of hydrophobicity consists of 3 + 3 + 15 = 21 features (Jiu-Xin Tan1, 2019).

G-gap dipeptide composition (g-gap)

Only the association between two neighboring amino acid residues will convey the adjacent dipeptide composition. Indeed, in three-dimensional space, amino acids with g-gap residues can be adjacent to them (H. Tang W. C., 2016). We used the g-gap dipeptide composition that extends from neighboring dipeptides to find significant similarities in protein sequences. By using this method, a protein P can be articulated as below.

$$\mathbf{P} = \left[v_1^g, v_2^g, \dots, v_i^g, \dots v_{400}^g\right]^T \quad (7)$$

Here the T symbol show the transposition of the vector; the v_i^g in this present the frequency of the i-th (i = 1,2,…,400) g-gap dipeptide and could be expressed as: $v_i^g = \frac{n_i^g}{L-g-1}$ (8)

Where n_i^g is the number present as the, i-th g-gap dipeptide; L is the measurement of the protein P; g is the

number of amino acid remains parted by two amino acid scums.

In this research, we reviewed the cases of g ranging "between" (1 to 9) as the case of g = 0 has been deliberate in reference (H. Tang W. C., 2016).

DEEP NEURAL NETWORK (DNN)

The deep neural network is a multilayers network architecture that is used for processing nodes. Whereas all the number of nodes are involved in the connected layers are used to learn and determine the hierarchical representation of data with an increasing level of abstraction (Shahid Akbar, 2020). DNN architecture consists of several layers such as the input layer, output layer, and hidden layers. Apart from the traditional neural network techniques, DNN is considered more significant due to more hidden layers However, the use of more hidden layers in a network has a high impact on the training capabilities and effectiveness of a model. In additionally, high-performance results of DNN over the conventional machine learning algorithms consider more efficient and reliable in term of training and testing (D. Ravì, 2017) DNN is widely applied by researchers in several areas such as speech recognition, image processing, biomedical engineering and natural language processing. In this job, along with the input and output layer, the DNN model is trained using six hidden layers. The feature vector X is first given to the input layer, while each node is mapped with the weight-based instance that calculates the output, is the bias vector, and denotes the activation function. The output values of the input layer are then given to the 1st hidden layer that uses the weight W2, bias vector B2 and activation function to measure the output. The process is continued and all the dense hidden layers are moved through before the output layer is reached. Where X represents the input vector, Y is the output vector, B is the bias vector, n is the nth layer activation function, n denotes the layer no, and W_n it represents the nth layer weight matrix. In our DNN model, two non-linear activation functions are used to change the learning rate, i.e., Tanh and soft-max (W.W. Fok, 2018). Whereas Tanh is used on hidden layers and on the output layer, soft-max is used.

Choosing optimal hyper-parameters has a major effect on the success rates of a classification model in deep learning algorithms. The highest performance accuracies of 94.11 percent and 92.31 percent were achieved using training and independent datasets, respectively, according to the results of our proposed model, by preserving the learning rate of 0.1. In addition, the activation function of Tanh and 500 no; was used to achieve higher results from training iterations. Moreover, it was also found that our proposed model's error loss was substantially reduced by increasing the number of iterations of training. We have used L1 regularization (LASSO i.e. Least Absolute Shrinkage and Selection Operator), L2 regularization (Ridge Regression), and dropped techniques to solve the over-fitting problem of the model (S. Khan, 2020)

GRADIENT BOOSTING (GB)

T he GB algorithm (Friedman, 2001,) a advancing learning ensemble method which produces a good ultimate prediction model, was proposed by Friedman. Centered on an ensemble of weak models (decision trees) commonly used in bioinformatics and biology of computation (Manavalan B, 2018.). In GB, ntree and nsplit are the two most important parameters, and we optimized the search space as described in (Manavalan B, 2018.). We note that there are other algorithms, in addition to the above algorithms of machine-learning (S. Basith, 2018). ML algorithms such as network of deep conviction, recurrent neural network, In different biological issues, deep learning and the two-layer neural network have been successfully applied (Cao R, 2017.).

POSITION SPECIFIC SCORING MATRIX (PSSM)

Several studies in the past have indicated the evolutionary characteristics. They have more discriminatory trends than their significant classifications. These techniques are therefore used extensively in the design of many predictive models, such as DNA (Deoxyribonucleic acid) binding for biological problems Protein detection, classification of membrane protein types and protein fold recognition (Shahid Akbar, 2020). The progression data is computed by a descriptor of a function known as Location Specific Scoring Matrix (the PSSM). The method was also known as the Position-Specific Weight Matrix (PSWM) (Shahid Akbar, 2020). PSSM has a sizeL*20 for protein arrangement with a length L, which Formulated through the PSI-BLAST method, the Swiss-Prot database is searched for. We used the alignment of multiple sequences (M. Waris, 2016,) with three iterations and 0.001 cut-off value. The $(m,n)t^h$ displaying a residue's ith position in the residue The question sequence that is mutated during the evolution with *jth* residue The continuing. The representative shape of a PSSM is the following P.

PSEUDO-POSITION SPECIFIC SCORING MATRIX (PPSSM)

PSSM, however, discovers evolutionary knowledge, but Machine-learning, regardless of the difference in the span of protein arrangements, it is difficult to deal explicitly with algorithms such as RF, KNN, and SVM (Shahid Akbar, 2020). In addition, PSSM ignores details about the sequence order and Factors for correlation. To keep up with these problems, we used Pse-PSSM to Find the data for sequence order and compute the frequencies each of the residues (K. C. Chou, 2007). In the arenas of bioinformatics and proteomics (Shahid Akbar, 2020), Pse-PSSM has various applications, including DNA binding proteins, prediction of protein structural classes, and

protein crystallization. The Pse-PSSM can be formulated by the below equation.

$P_{pse} = [p1; p2; \ldots; p20; p^{\varepsilon}{}_{1}; p^{\varepsilon}{}_{2}\ldots., p^{\varepsilon}{}_{20}]^{T}$

PSEUDO AMINO ACID COMPOSITION (PSE-AAC)
Not only can the amino acid composition, but also the association of physicochemical properties between two residues be included in the Pse-AAC method (Chou, 2011). We adopted the Type II Pse-AAC in this paper, in which a sample protein P can be formulated as follows (Jiu-Xin Tan1, 2019).

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

Where '9' is the number of physicochemical amino acid properties taken into account, namely, hydrophobicity, hydrophilicity, mass, pK1, pK2, pI, rigidity, flexibility and irreplaceability, 'λ 'is the correlation range;" is the frequencies for each factor and is formulated as:

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

Where ω is the weight factor for the impact of the sequence order; the frequency of the 400 dipeptides, the correlation factor between residues for the physicochemical properties will be as (τ_j). Additional comprehensive information on the process of deriving the formula could be found in reference (F. Y. Dao, 2017).
In this article, with the step of 1, the parameter λ is 1 to 95 and with the step of 1, the parameter ω is 0.1 to 1 with the step of 0.1 that is why, 10×95 = 950 function subsets based on Pse-AAC will therefore be obtained.

EXTREMELY *RANDOMIZED TREE (ERT)*
The ERT algorithm, which utilizes hundreds or thousands of independent decision trees to perform classification and regression problems, was introduced by (Geurts P, 2006.). It has been implemented in a large number of Problems with genetics. (S. Basith, 2018) The aim of ERT is to further reduce the variance of by including stronger randomisation methods, the prediction model. Similar to RF, the ERT algorithm is similar, but with the following differences. (S. Basith, 2018) (i) ERT shall not use the bagging technique for the construction of luggage. Every single tree. Instead, for the construction of each tree, it uses the whole input training collection. (ii) ERT selects a split node very randomly (both the index variable and the split variable values are randomly selected) (S. Basith, 2018), while among a random subset of variables, RF finds the best split (optimized by a variable index and a variable splitting value).

TRIPEPTIDE COMPOSITION (TPC)
Tripeptide is serene of three contiguous amino acids in a protein sequence, (Jiu-Xin Tan1, 2019) which is a bio signaling with least functionality. By espousing TPC, a sample protein P can be formulated by :

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

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

$$t_i = \frac{n_i}{L-2} \quad (12)$$

Where i-th tripeptide number is n_i ; L represent length of the protein P.

FEATURE SELECTION METHODS
ANALYSIS OF VARIANCE (ANOVA)
Feature selection is vital to develop the classification performance. It can filter the noisy features (Q. Zou, 2016). We approved the ANOVA technique to select optimal features from g-gap dipeptide compositions and PseAAC (Jiu-Xin Tan1, 2019). The ANOVA technique can measure the ratio of the modification among groups and the variance within groups for each attribute (Jiu-Xin Tan1, 2019). The formula expressions is defined as follows:

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

Where i-th feature score is $F(i)$, a high $F(i)$ value means a high ability to identify the sample; $S_w^2(i)$ is the modification within groups; $S_b^2(i)$ is the inconsistency among groups; and they may be planned as follows:

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

Where $SS_b(i)$ will be the sum of the squares among the groups; $SS_w(i)$ will be the sum of squares inside the

groups; K is the entire number of classes; N is the whole number of samples (Jiu-Xin Tan1, 2019).

BINOMIAL DISTRIBUTION (BD)

We approved the BD method to choose optimal features from tripeptide composition (Jiu-Xin Tan1, 2019). In this algorithm, the confidence level (CL) of every feature can be measured as:

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

While CL_{ij} is the confidence level for the i-th tripeptide in the j-th type; j signifies the kind 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. Rendering to the formula as distinct in Eq. (15), a high CL-value revenue a high ability to identify the sample. The BD method can abstract the over-represented keynotes, which is a brilliant statistical method broadly used in bioinformatics (Jiu-Xin Tan1, 2019).

INCREMENTAL FEATURE SELECTION (IFS) PROCESS

In general, it will not provide adequate information if a model is based on a low-dimensional function subset. On the conflicting, if a model is constructed on a high-bulk subset of structures, it can lead to duplication of information and overfitting issues. Therefore, the IFS method and 5-fold cross-validation of the ANOVA and BD method were used to examine the ideal feature set with optimum accuracy (Jiu-Xin Tan1, 2019) (Figure 3). According to the F(i)-values or CL-values, we rated all features and obtained new feature vectors, which are shown below.

$$\mathbf{P}' = [g_1', g_2', \dots g_n']^T \quad (16)$$

The initial feature subset contains the characteristic with the peak F(i)-value or CL-value,

$$\mathbf{P}' = [g_1']^T;$$

By accumulation the second peak F(i)-value or CL-value to the first subset, the second feature subset $\mathbf{P}' = [g_1', g_2']^T$ is shaped. The technique was repeated until all features were measured.

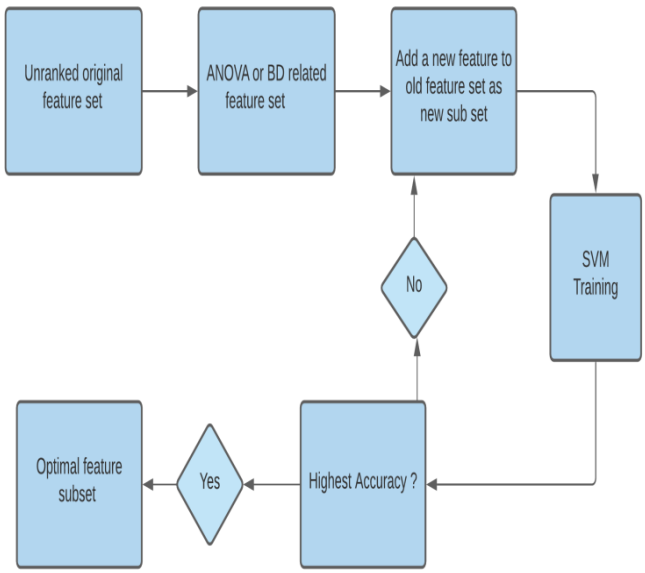

Figure 3. The background of the IFS process.

SUPPORT VECTOR MACHINE (SVM)

A supervised machine learning technique is the support vector machine (SVM) and has been commonly used in bioinformatics (W. Chen, 2018). Its key concept is to map the input features by nonlinear transformation from low-dimensional space to high-dimensional space and find the optimal linear classification surface. On behalf of convenience, you could download the LibSVM SVM software packages from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. In the current research, the package LibSVM-3.22 was adopted to investigate the efficiency of HBP identification (Jiu-Xin Tan1, 2019). In addition, to conduct predictions, the radical basis function kernel was chosen. The search spaces on the grid are $[2^{-5}, 2^{15}]$ with step 2 for penalty parameter C and $[2^3, 2^{-15}]$ for penalty parameter C with step of 2^{-1} for kernel parameter g.

PERFORMANCE EVALUATION

Three cross-validation methods are commonly used to investigate the efficiency of a predictor in realistic application, namely the independent dataset test, the sub-sampling test, and the jackknife test (W. Chen, 2018). The 5-fold cross-validation test to calculate the optimal parameters C and g of SVM in this paper was adopted in order to save computation time.

To test the models (Jiu-Xin Tan1, 2019), five assessment indexes were adopted (Renzhi Cao, 2017). Sensitivity (^{S}n) is used to test the ability of the model to predict positive samples correctly. Specificity (^{S}p) is used to test the ability of the model to predict negative samples correctly. The Overall Accuracy (**Acc**) represents the proportion that can

be correctly estimated for the entire benchmark dataset. The Matthew coefficient of correlation (**M**$_{CC}$) is used to determine the algorithm's reliability. The region under the ROC curve (AUC) represents the capacity of the model to distinguish decision values.

It is possible to compute them as follows:

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

(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.

I. RESULTS AND DISCUSSION

PERFORMANCES OF VARIOUS FEATURES

In this review and research, we observed the performance of 5 feature extraction methods and their combinations. Based on CTD, NV, CTD+NV methods, protein samples can be stated as 21-D (dimensional), 60-D and 81-D vector, (Jiu-Xin Tan1, 2019). The **A**$_{CC}$'s obtained was 60.16%, 70.33% and 67.07% by using SVM in to 5-fold cross-validation, (as shown in Table 1). It was originated that the estimate performances were far from satisfactory.

Centered 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 got 9 feature subsets. Firstly, we inspected the performances of these 400-D features subsets based on SVM. The results were reported in Figure 4A. Successively, the "ANOVA" method (Jiu-Xin Tan1, 2019) through the IFS process was applied to scrutinize the optimal technique and the results were recorded in Figure 4B. One may notice that while g = 1, a maximum **A**$_{CC}$ of 80.89% was obtained when the top 144 features were used. Obviously, **A**$_{CC}$'s were significantly increased by adopting ANOVA method. Although, prediction performances still needed to develop. In Pse-AAC based method, we attained 95×10 = 950 (95 kinds of λ and 10 kinds of ω) feature subsets. Initially, we examined the performances of these 950 models by using SVM (Jiu-Xin Tan1, 2019) in to 5-fold cross-validation test and reported the results in Figure 5A. It was originated that the maximum **A**$_{CC}$ of 76.83% was achieved when λ = 18 and ω = 0.1. In order to improve **A**$_{CC}$, the ANOVA method was adopted to rank the 400 + 18×9 = 572 features. By implementing SVM with IFS, a maximum **A**$_{CC}$ of 84.15% was acquired when the top 194 features were used (Figure 5B). Although the result was encouraging, the **A**$_{CC}$ 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	0.404	0.762
CTD+NV (81-D)	2^9	2^{-7}	70.73	63.41	67.07	0.342	0.762

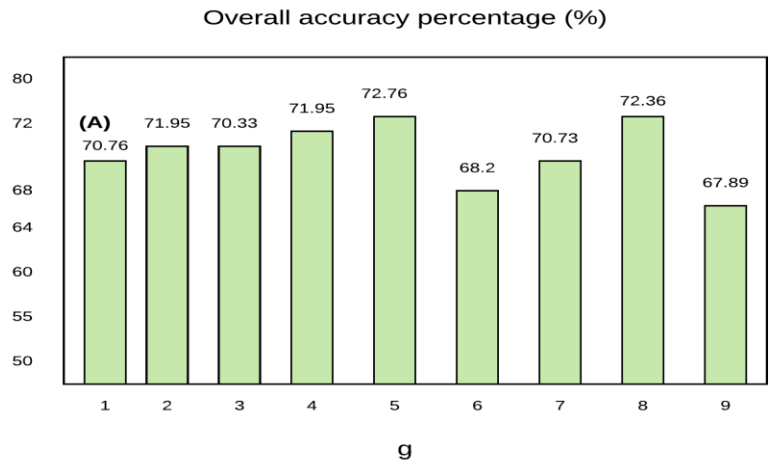

Figure 4. *Acc*'s for g-gap dipeptide composition. Different g values corresponding to different *Acc*'s.

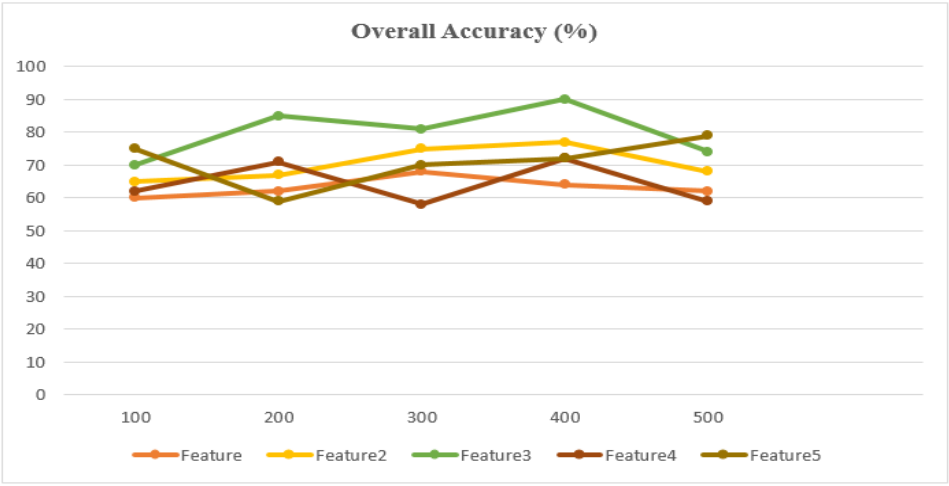

Figure 5. A plot showing the IFS curve based on Pse-AAC method.

Based on the TPC method, round about 8500 topographies were extracted for separately protein sequence. Meanwhile it would central 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 6). In this case, the S_n, S_p and *Mcc* are 96.75%, 97.56%, and 0.943, (Jiu-Xin Tan1, 2019). 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

.

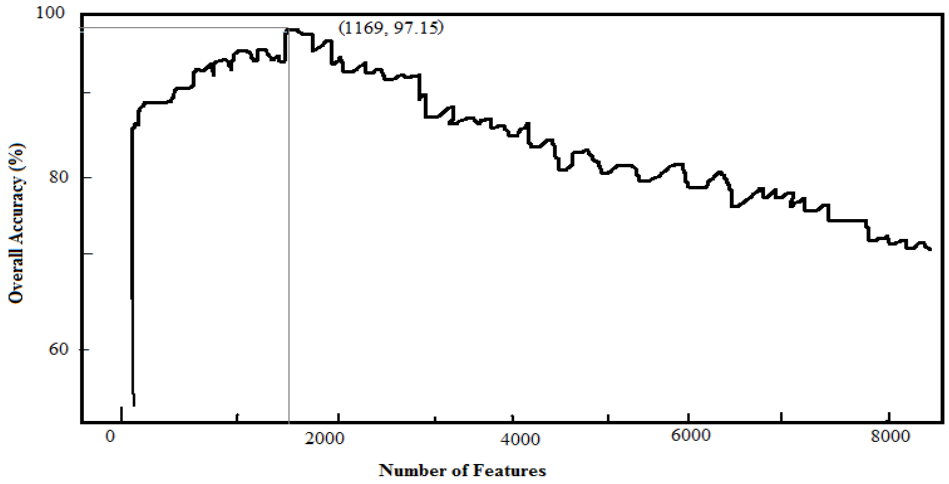

Figure 6. A plot showing IFS curvature based on TPC method.

COMPARISON WITH OTHER TECHNIQUES

In directive to display the dominance of SVM to identify HBP, we contrast its performance with folks of other machine learning algorithms (Jiu-Xin Tan1, 2019) based on the same feature subset (i.e. 117 optimal features). Id algorithm from the prediction of HIV proteins (Zhao, 2018.). From Table 2, we could find that the SVM classifier that will produce the best performance in these algorithms. Thus, the final model was constructed based on S

Table 2. Matching 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	69.12	0.392	0.770
Id Algorithm	82.23	64.14	72.15	0.7165	_
RT	84.15	69.23	74.32	0.720	_
KNN	85.75	75.18	80.47	0.613	_
ERT	84.23	77.64	80.32	0.547	0.814
GB	84.21	77.98	80.45	0.544	0.812
IBK	85.13	71.35	79.45	0.725	_
DNN	86.11	85.51	85.21	0.704	_
Random Forest	88.62	84.56	86.59	0.734	0.947
Naive Bayes	95.93	92.68	94.31	0.887	0.965
SVM	96.75	97.56	97.15	0.943	0.994

Classifier	S_n(%)	S_p(%)	*Acc*(%)	*Mcc*	AUC
J48	63.41	56.91	60.16	0.204	0.601
Bagging	80.49	57.72	69.12	0.392	0.770
Id Algorithm	82.23	64.14	72.15	0.7165	_
RT	84.15	69.23	74.32	0.720	_
KNN	85.75	75.18	80.47	0.613	_
ERT	84.23	77.64	80.32	0.547	0.814
GB	84.21	77.98	80.45	0.544	0.812
IBK	85.13	71.35	79.45	0.725	_
DNN	86.11	85.51	85.21	0.704	_
Random Forest	88.62	84.56	86.59	0.734	0.947
Naive Bayes	95.93	92.68	94.31	0.887	0.965
SVM	96.75	97.56	97.15	0.943	0.994

The first column represents the method name developed in this study. The second, third, fourth, fifth, and sixth columns, respectively, represent the sensitivity, specificity, accuracy, MCC and AUC. RF: random forest; KNN: k nearest neighbor; ERT: extra tree classifier; GB: gradient boosting; DNN: Deep Neural Network and SVM: support vector machine. So, it is also essential to compare the chosen technique suggested in this review paper with existing techniques. **Table 3** displays the detailed results of different techniques for identifying. HBP. Based on the same benchmark dataset, (H. Tang Y. W., 2018) achieved an accuracy ***Acc*** of 84.9% by using a SVM-based method, where proteins sequences were determined using the optimal 0-gap dipeptide composition features obtained by (H. Tang Y. W., 2018) the ANOVA feature selection technique. (Q. Zou, 2016) Achieved the accuracy of 89 percent by using the method of TBT. (S. Khan, 2020)Achieved the accuracy of 84.60 percent by using the method of GA-WE. (Adele Cutler, 2001,)Achieved the accuracy of 88.72 percent by using the method of RF. (K. Wang, (2019)Achieved the accuracy of 90.71 percent by using the method of Ensemble. (S. Khan, 2020)Achieved the accuracy of 89.60 percent by using the method of piRNAPred. (S. Basith, 2018) Obtained a (CC) correlation coefficient and accuracy ($\boldsymbol{A_{CC}}$) of 84.96% in cross-validation experiment by training an extremely randomized tree with optimal features obtained from dipeptide composition and amino acid index values based on two-step feature selection. Overall performance of (Shahid Akbar, 2020) was better by using DeepPSSM he got good accuracy of 94.41%. Our proposed method (Jiu-Xin Tan1, 2019) could produce an accuracy ($\boldsymbol{A_{CC}}$) 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 selected method with further published method.

Reference	**Technique**	$\boldsymbol{S_n}$**(%)**	$\boldsymbol{S_p}$**(%)**	***Acc*(%)**	***Mcc***	**AUC**
(H. Tang Y. W., 2018)	HBPred	88.6	81.3	84.9	_	_

(S. Khan, 2020)	GA-WE	90.10	78.10	84.60	0.691	_
(S. Basith, 2018)	iGHBP	88.62	81.30	84.96	_	0.701
(Q. Zou, 2016)	TBT	89.60	91.10	90.46	_	_
(S. Khan, 2020)	piRNAPred	90.10	87.10	89.60	0.761	_
(S. Basith, 2018)	ERT	80.72	83.9	82.3	0.646	0.813
(Adele Cutler, 2001,)	Random Forest (RF)	87.18	75.14	88.72	0.472	0.777
(K. Wang, (2019)	Ensemble	92.71	87.92	90.71	_	_
(Shahid Akbar, 2020)	DeepPSSM	96.32	94.12	94.41	0.88	_
(Jiu-Xin Tan1, 2019)	HBPred2.0	96.75	97.56	97.15	0.943	0.994

For deep measure the performance of these methods, an independent dataset was used. The results were noted in Table 4. One may observe that the HBPred2.0 predictor achieved the best performance in three predictors, suggesting that HBPre2.0 has well generalization ability.

Table 4. Performance evaluation based on the independent dataset.

Reference	**Technique**	S_n(%)	S_p(%)	*Acc*(%)	*Mcc*	AUC
(H. Tang Y. W., 2018)	HBPred	80.43	56.52	68.48	0.381	0.715
(S. Basith, 2018)	Extra Tree Classifier	80.72	83.9	82.3	0.646	0.813
(Friedman, 2001,)	GB	77.12	54.8	66.14	0.331	0..700
(Khaled Fawagreh, 2014,)	Random Forest	86.52	83.19	85.01	0.691	
(S. Khan, 2020)	piRNA(2L)-PseKNC	82.60	83.91	83.54	0.691	_
(Adele Cutler, 2001,)	Random Forest (RF)	87.18	87.18	88.72	0.472	0.777
(D. Cheng, 2014)	KNN	73.01	82.51	78.23	0.598	_
(S. Basith, 2018)	iGHBP	86.96	47.83	67.39	0.380	_
(Shahid Akbar, 2020)	DeepPSSM	88.11	79.23	82.25	0.612	_
(K. Wang, (2019)	Ensemble	82.72	79.32	81.14	_	_
(Q. Zou, 2016)	TBT	87.12	78.12	81.75	_	_
(Jiu-Xin Tan1, 2019)	HBPred2.0	89.13	80.43	84.78	0.698	0.814

Specificity could reflect the discriminated competence of model arranged negative samples. From above Table 4, a higher specificity of the HBPred2.0 indicates that the model could produce less false positives (Jiu-Xin Tan1, 2019).

II. CONCLUSION

In this research, we systematically investigated the performances of numerous techniques and classifiers on HBP prediction. By a great number of experiments, we attained an accurate, reliable and effective the best model by combining SVM with optimal tripeptide composition to identify the HBP's. This model could produce the overall accuracy of 84.78% on the independent data. Finally, Due to published database (Zhi-Yong Liang, 2016) and webserver (Jiangning Song, 2019) could provide more convenience for scientific community. They established a free webserver for the proposed method "HBPred2.0" (Jiu-Xin Tan1, 2019) which can be free accessed form http://lin-group.cn/server/HBPred2.0/. We expect that the tool will help scholars to identify the HBP'S and study the mechanism of HBP's function and promote the development of related drug research. Consequently, we hope it will be considered a useful tool for the research community.

References

1. Adele Cutler, B. L. (2001,). Random forests, . "Mach Learning,", 1-19.
2. Cao R, A. B. (2017.). QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. . Bioinformatics, 33. , 586–598.
3. Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273, 236-247.
4. D. Cheng, S. Z. (2014). kNN Algorithm with Data-Driven k Value,. "Adv. Data Min. Appl.,", 499-512.
5. D. Ravì, C. W.-P.-Z. (2017). Deep learning for health informatics, . IEEE journal of biomedical and health informatics, 4-21.
6. F. Y. Dao, H. Y. (2017). Recent advances in conotoxin classification by using machine learning methods. Molecules, 22, 1-21.
7. Friedman, J. H. (2001,). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, Vol. 29,, 1189-1232.
8. G. Bauman. (2002,). "Growth hormone binding protein. The soluble growth hormone receptor,". Minerva Endocrinol. 27,, 265–276.
9. G. Ozzola. (2016,). Essay of sex hormone binding protein in internal medicine: a brief review,. La Clinica Terapeutica 167, 127-129.
10. Geurts P, E. D. (2006.). Extremely randomized trees. . Extremely randomized trees. Mach Learning 2006,.
11. H. Tang, W. C. (2016). Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Molecular Biosystems, 1269-1275.
12. H. Tang, Y. W. (2018). HBPred: a tool to identify growth hormone-binding proteins, Int. J. Biol. Sci. HBPred: a tool to identify growth hormone-binding proteins, Int. J. Biol. Sci., 957–964.
13. I.E. Einarsdottir, N. G. (2014). Plasma growth hormone-binding protein levels in Atlantic salmon Salmo salar during smoltification and seawater transfer. Journal of Fish Biology (2014) 85, 12791296.
14. Jiangning Song, Y. W. (2019). "iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites". Briefing in Bioinformatics, 20(2), 2019.,, 638-658.
15. Jiu-Xin Tan1, S.-H. L.-M.-X. (2019). Identification of hormone binding proteins based on machine learning methods . Identification of hormone binding proteins based on machine learning methods , 2467.
16. K. C. Chou, H.-B. S. (2007). MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM,. Biochem. Biophys. Res. Commun. 360.,, 339-345.
17. K. Wang, S. L. ((2019). " Identification of hormone-binding proteins using a novel ensemble classifier,". Computing 101.,, 693–703.
18. Khaled Fawagreh, M. M. (2014,). Random forests: from early developments to recent advancements. "Systems Science & Control Engineering,", 1-9.
19. L. Breuza, S. P. (2016). The UniProtKB guide to the human proteome. Database (Oxford), 2016 , 1-10.
20. L. Fu, B. N. (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data,". Bioinformatics, 28 (2012),, 3150–3152.
21. L., B. (2001,). Random forests. Random forests. Mach Learning ,45., 5-32.
22. M. Waris, K. A. (2016,). Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix,. Neurocomputing 199., 154-162.
23. Madias, J. A. (2017). Adverse Effects of the Metabolic Acidosis of Chronic Kidney Disease. Adverse Effects of the Metabolic Acidosis of Chronic Kidney Disease, 289-297.
24. Manavalan B, G. R. (2018.). iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. Front Immunol 2018;9:1695.
25. Q. Zou, S. W. (2016). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy . BMC System Biology, 401-412.
26. Renzhi Cao, C. F. (2017). "ProLanGO: Protein function prediction using neural machine translation based on a recurrent neural network". Molecules 2017, 22, 1732, 1-14.
27. S. Basith, B. M. (2018). iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Computational and Structural Biotechnology Journal 16, 412-420.
28. S. Khan, M. K.-C. (2020). Prediction of piRNAs and Their Function Based on Discriminative Intelligent Model Using Hybrid Features into Chou's PseKNC, Chemometrics and Intelligent Laboratory Systems. Prediction of piRNAs and Their Function Based on Discriminative Intelligent Model

Using Hybrid Features into Chou's PseKNC, Chemometrics and Intelligent Laboratory Systems.
29. Shahid Akbar, S. K. (2020). iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach , . iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach , , 1-11.
30. W. Chen, P. F. (2018). Recent advances in machine learning methods for predicting heat shock proteins. Drug Metab in press.
31. W.W. Fok, Y. H. (2018). Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine. 4th International Conference on Information Management (ICIM), 103-106.
32. Y. Zhang, T. M. (1999). Identification of serum GH-binding proteins in the goldfish (Carassius auratus) and comparison with mammalian GH-binding proteins. J. Endocrinol, GH-binding proteins, 255-262.
33. Zhao, J. M. (2018.). Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers.
34. Zhi-Yong Liang, H. L.-J. (2016). "Pro54DB: a database for experimentally verified sigma-54 promoters". Bioinformatics, 33(3), 2017, 467-469.
35. Saeed, S.; Mahmood, M. K.; Khan, Y. D., An exposition of facial expression recognition techniques. Neural Computing and Applications 2018, 29 (9), 425-443.
36. Butt, A. H.; Khan, Y. D., CanLect-Pred: A cancer therapeutics tool for prediction of target cancerlectins using experiential annotated proteomic sequences. IEEE Access 2019, 8, 9520-9531.
37. Amanat, S.; Ashraf, A.; Hussain, W.; Rasool, N.; Khan, Y. D., Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC. Current Bioinformatics 2020, 15 (5), 396-407.
38. Ilyas, S., Hussain, W., Ashraf, A., Khan, Y. D., Khan, S. A., & Chou, K. C. (2019). iMethylK-PseAAC: Improving accuracy of lysine methylation sites identification by incorporating statistical moments and position relative features into general PseAAC via Chou's 5-steps rule. Current Genomics, 20(4), 275-292.
39. Hussain, W.; Rasool, N.; Khan, Y. D., A Sequence-Based Predictor of Zika Virus Proteins Developed by Integration of PseAAC and Statistical Moments. Combinatorial chemistry & high throughput screening 2020, 23 (8), 797-804.
40. Khan, Y. D.; Alzahrani, E.; Alghamdi, W.; Ullah, M. Z., Sequence-based Identification of Allergen Proteins Developed by Integration of PseAAC and Statistical Moments via 5-Step Rule. Current Bioinformatics 2020, 15 (9), 1046-1055.
41. Mahmood, M. K.; Ehsan, A.; Khan, Y. D.; Chou, K.-C., iHyd-LysSite (EPSV): Identifying Hydroxylysine Sites in Protein Using Statistical Formulation by Extracting Enhanced Position and Sequence Variant Feature Technique. Current Genomics 2020, 21 (7), 536-545.
42. Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., IPhosS (Deep)-PseAAC: Identify phosphoserine sites in proteins using deep learning on general pseudo amino acid compositions via modified 5-Steps rule. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2020.
43. Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC. Current Bioinformatics 2020, 15 (8), 937-948.
44. Shah, A. A.; Khan, Y. D., Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. Scientific Reports 2020, 10 (1), 1-10.
45. Awais, M.; Hussain, W.; Rasool, N.; Khan, Y. D., iTSP-PseAAC: Identifying Tumor Suppressor Proteins by Using Fully Connected Neural Network and PseAAC. Current Bioinformatics 2021, 16 (5), 700-709.
46. Hussain, W.; Rasool, N.; Khan, Y. D., Insights into Machine Learning-based approaches for Virtual Screening in Drug Discovery: Existing strategies and streamlining through FP-CADD. Current Drug Discovery Technologies 2021, 18 (4), 463-472.
47. Khan, Y. D.; Khan, N. S.; Naseer, S.; Butt, A. H., iSUMOK-PseAAC: prediction of lysine sumoylation sites using statistical moments and Chou's PseAAC. PeerJ 2021, 9, e11581.
48. Malebary, S. J.; Khan, R.; Khan, Y. D., ProtoPred: Advancing Oncological Research Through Identification of Proto-Oncogene Proteins. IEEE Access 2021, 9, 68788-68797.
49. Malebary, S. J.; Khan, Y. D., Evaluating machine learning methodologies for identification of cancer driver genes. Scientific reports 2021, 11 (1), 1-13.
50. Malebary, S. J.; Khan, Y. D., Identification of Antimicrobial Peptides Using Chou's 5 Step Rule. CMC-COMPUTERS MATERIALS & CONTINUA 2021, 67 (3), 2863-2881.
51. Naseer, S.; Ali, R. F.; Khan, Y. D.; Dominic, P., iGluK-Deep: computational identification of lysine glutarylation sites using deep neural networks with general pseudo amino acid compositions. Journal of Biomolecular Structure and Dynamics 2021, 1-14.
52. Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., NPalmitoylDeep-PseAAC: A Predictor of N-Palmitoylation Sites in Proteins Using Deep Representations of Proteins and PseAAC via Modified 5-Steps Rule. Current Bioinformatics 2021, 16 (2), 294-305.
53. Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. Analytical Biochemistry 2021, 615, 114069.
54. Khanum, S., Ashraf, M. A., Karim, A., Shoaib, B., Khan, M. A., Naqvi, R. A., ... & Alswaitti, M. Gly-LysPred: Identification of Lysine Glycation Sites in Protein Using Position Relative Features and Statistical Moments via Chou's 5 Step Rule.
55. Lv, H., Dao, F. Y., Zhang, D., Yang, H., & Lin, H. (2021). Advances in mapping the epigenetic modifications of 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC). Biotechnology and Bioengineering.

56. Zulfiqar, H., Sun, Z. J., Huang, Q. L., Yuan, S. S., Lv, H., Dao, F. Y., ... & Li, Y. W. (2021). Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods.
57. Liu, Y., Wang, X., & Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Briefings in bioinformatics, 20(1), 330-346.
58. Zhang, D., Xu, Z. C., Su, W., Yang, Y. H., Lv, H., Yang, H., & Lin, H. (2021). iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics, 37(2), 171-177.