Chemometrics and Intelligent Laboratory Systems 196 (2020) 103912

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

cACP: Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components

Shahid Akbar [a], Ateeq Ur Rahman [a], Maqsood Hayat [a,*], Mohammad Sohail [b]

[a] Department of Computer Science, Abdul Wali Khan University Mardan, KP, 23200, Pakistan
[b] Department of Physics, University of Lahore, Sargodha Campus, 40100, Sargodha, Pakistan

ARTICLE INFO

Keywords:
Anticancer peptides
QSO
SVM
RF
PCA

ABSTRACT

World widely, cancer is considered a fatal disease and remains the major cause of death. Conventional medication approaches using therapies and anticancer drugs are deemed ineffective due to its high cost and harmful impacts on the normal cells. However, the innovation of anticancer peptides (ACPs) provides an effective way how to deals with cancer affected cells. Due to the rapid increases in peptide sequences, truly characterization of ACPs has become a challenging task for investigators.

In this paper, an effort has been carried out to develop a reliable and intelligent computational method for the accurate discrimination of anticancer peptides. Three statistical feature representation schemes namely: Quasi-sequence order (QSO), conjoint triad feature, and Geary autocorrelation descriptor are applied to express motif of the target class. In order to eradicate irrelevant and noisy features, while select salient, profound and high variated features, principal component analysis is employed. Furthermore, the diverse nature of learning algorithms is utilized in order to select the best operational engine for the proposed model. After examining the empirical outcomes, support vector machine obtained quite encouraging results in combination with QSO feature space. It has achieved an accuracy of 96.91% and 89.54% using the main dataset and alternative dataset, respectively. It is observed that our proposed model shows an outstanding improvement compared to literature methods. It is expected that the developed model may be played a useful role in research academia as well as proteomics and drug development.

1. Introduction

Cancer is the most devastating disease and considered the major reason of death both in economically developed and undeveloped countries. Every year, about eight million people died from this lethal disease [1]. It is also estimated that till 2020, the ratio of cancer infected cases will be increased to 16 million [2,3]. Cancer treatment via conventional methods i.e., chemotherapy, radiation therapy, hormonal therapy, and targeted therapy are deemed unsuccessful due to the high cost and its harmful impacts on the normal cells [4,5].

Over the last few decades, anticancer peptides (ACPs) have been considered the most operative cancer treatment because it does not affect normal body physiological functions. It is pre-clinically used for different purposes, such as diabetes, cardiovascular diseases, and various types of tumors [6,7]. ACPs have exceptional and unique benefits such as high efficiency and safer than synthetic drugs [4]. ACPs have basically small peptides, where sequence size ranges from 5 to 30 amino acids residues. ACPs, due to its cationic nature, can easily treat cancer-affected cells [8]. However, the respective affected cells are eradicated by interacting with anionic cell membrane components of a cancer cell.

Identification of novel ACPs using experimental procedures is time-consuming and expensive. Due to the indispensable role the ACPs, the researchers and drug developers have adopted the concept of automation as an alternative tool for effective identification of ACPs.

In this regards, various automated intelligent models have been carried out by the investigators for the prediction of ACPs. Chen et al., proposed a model, called "iACP", for identification of anticancer peptides [9]. They have used an optimized G-Gap dipeptide composition is incorporated with formulating peptides sequences. Similarly; Manavalan et al., proposed a hybrid approach to predict ACPs [10]. Whereas, the hybrid feature vector consists of an optimal feature set of physiochemical properties, atomic and dipeptide composition. The proposed model is trained and tested using k-fold cross-validation. Moreover; Tyagi et al., developed in Silico models for prediction of ACPs [11]. The proposed

* Corresponding author.
E-mail address: m.hayat@awkum.edu.pk (M. Hayat).

https://doi.org/10.1016/j.chemolab.2019.103912
Received 23 March 2019; Received in revised form 6 May 2019; Accepted 18 December 2019
Available online 20 December 2019

model is measured using four different datasets. Whereas, the peptide sequences are formulated using binary profile and split amino acid composition. On the other hand, Li et al., developed hybrid features based predictor for the identification of ACPs [12]. Reduce amino acid composition (RAAC), average chemical shifts and amino acid composition (AAC) is applied to extract features. The proposed predictor reported an improved performance accuracy using SVM. Akbar et al., developed a novel model called "iACP-GAEnsC", for identification of anticancer peptides [13]. In this model, a composite encoding scheme is used to collect high discriminative features from peptide sequences. The performance outcomes of the developed approach are measured using an evolutionary genetic algorithm. Recently, Kabir et al., have developed "TargetACP", a novel intelligent predictor based on evolutionary and sequential information [14]. Synthetic Minority Over-sampling Technique (SMOTE) is applied to balance the samples among minority and majority classes. The proposed methodology was evaluated through two diverse benchmark datasets and obtained improved performance results. Furthermore; Vijayakumar et al., Presented a web server namely, "ACPP" for the design and accurate prediction of ACPs [15]. ACPP reported several modes, which permit the user to design and identify ACP more accurately. It also evaluates the query sequence whether it is having the apoptotic function or not. Similarly; Hajisharifi et al., have used pseudo amino acid composition and novel local alignment kernel to predict ACPs [16]. In a sequel, recently, Xu et al., have utilized g-gap dipeptide composition method as sequence representation [17]. The developed method has used maximum relevance-maximum distance (MRMD) to prune irrelevant and redundant features from feature space.

Series of current papers have demonstrated that [18–27], developing a new sequence-analyzing method or statistical predictor by observing the guidelines of Chou's 5-step rules: (i) Valid benchmark dataset selection or designing is the first step of developing predictor; (ii) the second step is to formulate the samples in such a way that can accurately reflect their internal correlation with the target to be identified; (iii) the third step is the selection of best operational engine; (iv) the fourth step is to evaluate the model via cross-validation tests; (v) and finally, launching a user-friendly and publically accessible web-server for the predictor.

In this paper, we have developed an effective and computational intelligent model for prediction of anticancer peptides. Three distinct protein samples formulation methods such as Quasi-sequence order (QSO), conjoint triad feature (CTF) and Geary Autocorrelation Descriptor are used to express peptide sequences. Furthermore, principal component analysis (PCA) is adopted to select high discriminative features from extracted feature spaces. Finally, three different classification learners such as SVM, KNN, and RF are utilized as operational engines to examine predictive outcomes of the proposed model.

The remaining paper is organized as follow: section 2 represents materials and methods, performance criteria are presented in section 3; results and discussions are discussed in section 4; at the end of the paper, the conclusion has been drawn.

2. Materials and methods

2.1. Dataset

Selection or construction of a valid dataset is an important step for statistical predictor because it has a great impact on classification measures. Keeping the significance of dataset, two different datasets i.e., main dataset and alternate dataset are applied [11]. However, both the datasets are categorized into two predefined classes such as anticancer and non-anticancer peptides. Whereas, the anticancer peptide sequences are selected from anuran defense peptides database (DADP) [28], antimicrobial database and peptides (APD, CAMP) [29,30]. Similarly; non-anticancer sequences are gathered by the random selection of peptides from Swiss-Prot proteins database [31]. Furthermore, AMPs are also obtained from the above-mentioned databases such as DADP, CAMP, and ADP. In case of the main dataset, the total number of sequences is 2475, in which 225 positive samples are anticancer and the remaining 2250 negative samples are considered as non-anticancer peptides. On the other hand, the alternate dataset composed of 225 anticancer peptides and 1372 non-anticancer peptides [11].

2.2. Feature extraction techniques

With the enormous increase of biological sequences in the post-genomic era, the main challenging jobs are how to formulate a biological sequence with a discrete model that retains substantial sequence-order information or major motif characteristic. Machine-learning algorithms (such as "Optimization" algorithm [18], "Covariance Discriminant" algorithm [19,20], "Nearest Neighbor" [21], and "SVM" algorithm [22,23]) can only use vectors as demonstrated in a comprehensive review [32]. The main issue reported in the discrete model is the loss of sequence-motif information. Pseudo amino acid composition (PseAAC) was introduced by Chou in order to preserve sequence-motif information for proteins, [24]. The concept of PseAAC has been extensively applied about all the areas of computational proteomics [26,27,31]. Four powerful open-access software, called 'PseAAC' [33], 'PseAAC-Builder' [34], 'propy' [35], and 'PseAAC-General' [36], were also established on the basis of PseAAC: the first three are for generating various modes of Chou's special PseAAC [37]; while the last one for those of Chou's general PseAAC [38], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode, "Gene Ontology" mode, and "PSSM" mode. After Encouraging successes of PseAAC in the area of proteins, it was extended to DNA\RNA using the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [39–41]. Recently a very powerful web-server called 'Pse-in-One' [42] and its updated version 'Pse-in-One2.0' [43] have been established.

In feature extraction step the input data are transformed into numerical descriptors, which represent the pattern of the target class [44,45]. Various feature representation approaches have been proposed in the literature to extract high discriminative feature in order to accurately identify protein sequences [46–48]. In this model, three different feature representation techniques, namely: Quasi sequence order (QSO), Geary autocorrelation and Conjoint triad feature (CTF) are effectively utilized to gather reliable, profound, and salient features from anticancer sequences.

2.2.1. Quasi-sequence order (QSO)

Owing to the huge number of sequence order patterns, it is hard for the developers to directly include the effect of sequence order into a statistical predictor [49,50]. Therefore, Quasi-sequence order (QSO) is used to indirectly incorporate the sequence order features. QSO descriptors are derived using both Grantham distance matrix and Schneider-Wrede distance matrix among each pair of 20 native amino acids [51,52]. Wherever Grantham matrix contains chemical distance information and Schneider-Wrede matrix computes the physicochemical properties i.e. hydrophilicity, hydrophobicity, side-chain volume and polarity [53].

Suppose a peptide chain of N amino acid residues is represented as:

$$R_1R_2R_3R_4R_5.........R_N \tag{1}$$

The QSO effect uses the set of sequence-order-coupling numbers to collect statistical information's from peptides sequences are represents as below:

$$\begin{cases} \tau_1 = \dfrac{1}{Q-1}\displaystyle\sum_{i=1}^{Q-1} J_{i,i+1} \\ \tau_2 = \dfrac{1}{Q-2}\displaystyle\sum_{i=1}^{Q-2} J_{i,i+2} \ , (\phi < Q) \\ \tau_\phi = \dfrac{1}{Q-\phi}\displaystyle\sum_{i=1}^{Q-\phi} J_{i,i+\phi} \end{cases} \tag{2}$$

where τ_1, represents the first rank sequence order coupling number (SOCN) [54], that reproduces the coupling among all most connected residue along a peptide sequence, τ_2 is the second SOCN among the second most adjacent residues, and so forth.

$$J_{i,j} = D^2\left(R_i, R_j\right) \tag{3}$$

where $J_{i,j}$ is the coupling factor of R_i, R_j and $D(R_i, R_j)$ is the physiochemical distance from R_i to R_j.

For each amino acid type, the QSO descriptors of a protein sequence q can be represented as:

$$q = \left[q_1, q_2, \ldots, q_{20}, q_{21}, \ldots, q_{20+nlag}\right] \tag{4}$$

$$\begin{cases} \dfrac{N_r}{\sum\limits_{i-1}^{20} N_r + w\sum\limits_{d=1}^{nlag} \tau_d}, & i-1,2,3,.....,20 \\ \dfrac{w\tau_{r-20}}{\sum\limits_{i-1}^{20} N_r + w\sum\limits_{i=1}^{nlag} \tau_d}, & i = 21,22,23,....20+nlag \end{cases} \tag{5}$$

where N_r represents the normalized occurrence of amino acid ‘i’ and ‘w’ denotes the weight factor.

2.2.2. *Conjoint triad feature (CTF)*

Conjoint triad feature (CTF) was initially proposed by shen et al. [55], to accurately represent the 20 native amino acids and to suit synonymous mutation [56]. Firstly, CTF divides the 20 amino acids into seven classes such as; [(H, N, Q, W), (A, G, V), (Y, M, T, S), (I, L, F, P), (D, E), (R, K), (C) [57,58]. The clustering of the amino acids into seven classes is based on volumes and dipoles of slide chains [59]. Amino acids of the similar class contain the synonymous mutation due to their identical properties [60]. CTF expresses the biological sequences by computing the frequency of each triad type [61], where triad is a unit of continuous three amino acids belongs to the same class and are treated identically [62]. Finally, the resultant feature vector having dimensions of n*343 is obtained, where n represents the number of sequences and 343 are features against each protein sequence. CTF is a composite features space of amino acid composition (AAC) and sequence order information.

A peptide sequence R of N amino acid residues are represented as:

$$R_1R_2R_3R_4R_5.........R_N \tag{6}$$

Let us consider a binary vector (V, F) to express a peptide sequence, where V is the feature space of the extracted information’s, F is the frequency vector corresponding to V. However, the value of $\mathbf{f}_i$ correlates to the length of a protein sequence. Generally, a longer peptide sequence may have a larger value of f_i, which complicates the comparison among two heterogeneous peptides. To solve this issue, a parameter d_i is formed by normalizing the frequency f_i, of each peptide sequence represented as below:

$$d_i = \frac{f_i - \min\left\{f_0 f_2, \ldots, f_{342}\right\}}{\max\left\{f_0 f_2, \ldots, f_{342}\right\}} \tag{7}$$

2.2.3. *Geary Autocorrelation Descriptor*

Geary autocorrelation is an efficient statistical approach [63], that uses the composition of the amino acid property along time sequence, which basically calculates the correlation among between two residues separated by a lag distance in terms of their evolution measures [64]. Geary calculates the spatial autocorrelation, which represents the correlation of a variable with itself through space [65].

Geary autocorrelation descriptors for peptide sequence can be represented as:

$$Geary(d) = \frac{\frac{1}{2(N-d)} \sum\limits_{i=1}^{N-d} \left(P_i - P_{i+d}\right)^2}{\frac{1}{(N-1)} \sum\limits_{i=1}^{N} \left(P_i - P'\ \right)^2} \quad d = 1,2,3,......... \tag{8}$$

where d denotes the log value of autocorrelation and P_i and P_{i+d} represents the properties of amino acids at location i and $i+d$, respectively.

2.3. *Feature selection*

In machine learning, the extracted feature vector is highly important and effectively used for the prediction of biological datasets. However, the feature spaces with high dimensionality may lead to erroneous and ineffective classification results [66]. Moreover, it also requires high computation time and memory to train and test a proposed model [67]. In order to overcome these issues, various feature selection schemes have been utilized in order to reduce the feature space [68–70]. Feature selection is the process to minimize the redundant and irrelevant features and improve the prediction result. In this model, Principal Component Analysis (PCA) is applied for the feature selection purpose. PCA transforms the number of correlated attributes into the small number of uncorrelated attributes [71]. The computed uncorrelated variables are known as principal components. The main advantages of PCA are to decrease the dimensionality of a feature vector with minimum correlation and minimum loss of discriminative features [72]. The global Euclidean structure of PCA makes it more sensitive to the outliers [73].

Let us consider, a feature vector ‘Y’, having the dimensions of P*Q, where “P” is the number of extracted features, “Q” is the number of peptide sequences, and “K" represents the required dimension of the feature vector. The value of K must be smaller than “Q”. In order to reduce the dimensionality, PCA uses the following steps:

1. Calculate the means of each attribute:

$$\overline{x_j} = \frac{1}{P}\sum_{i=1}^{P} x_i \tag{9}$$

2. The difference the mean x from $\mathrm{x_i}$:

$$\delta_i = x_i - \overline{x} \tag{10}$$

3. Compute the covariance matrix:

$$C_m = (x_i - \overline{x})(x_i - \overline{x})^T BB^T \tag{11}$$

where

$$B = \left\{\delta_1, \delta_2, ..., \delta_p\right\}(Q * P) \tag{12}$$

4. Calculate the eigenvalues of C_m:

$$\alpha_1 > \alpha_2, \ldots.., \alpha_N \tag{13}$$

where the highest eigenvalues “α_1“must be less than the second highest eigenvalues “α_2” and so on.

5. Calculate the eigenvector of

$$C_m : V_1, V_2, \ldots., V_N \tag{14}$$

6. Finally, select the require “k”, having the highest eigenvalues.

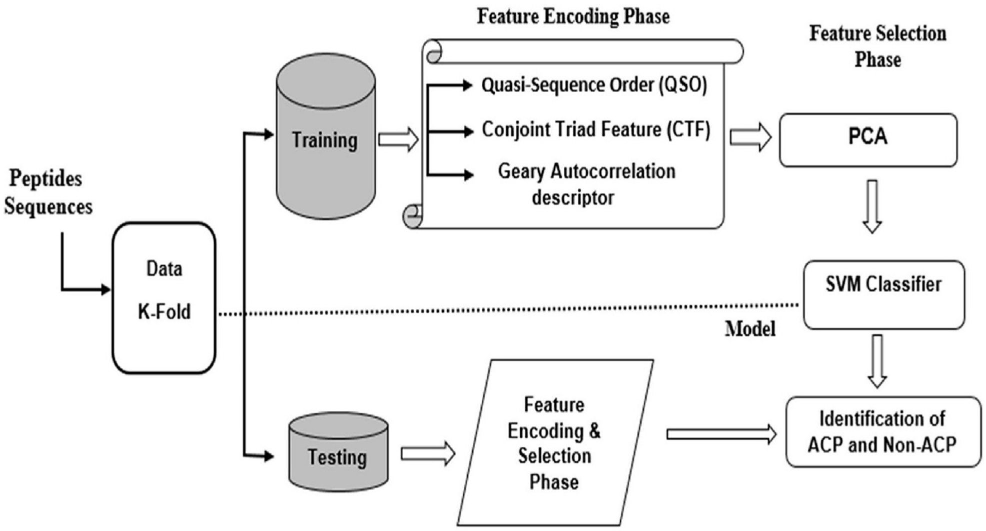
Feature Encoding Phase
Feature Selection Phase
Training
Quasi-Sequence Order (QSO)
Conjoint Triad Feature (CTF)
Geary Autocorrelation descriptor
PCA
Peptides Sequences
Data
K-Fold
SVM Classifier
Model
Testing
Feature Encoding & Selection Phase
Identification of ACP and Non-ACP

Fig. 1. Framework of our Proposed Model.

2.4. Classification techniques

2.4.1. Support vector machine (SVM)

Support vector machine is a widely used supervised learning algorithm that was initially proposed by Vapnik in 1995 for the binary class problem [74]. SVM performs efficiently for both linear and non-linear surfaces due to its effective training procedure [75]. SVM uses the statistical learning theory and was initially applied for only binary class problems but later on, it was extended to multiclass problems [76]. In case of the binary problem, SVM converts data into a high dimension feature vector to compute an optimum hyperplane. SVM uses various kernel functions such as linear, polynomial, Radial base function (RBF) and sigmoid in order to measure classification power. In the current study, RBF kernel function was used, whereas two parameters; γ and C to examine the benchmark datasets. The values of C and γ are determined via an optimization procedure using a grid search approach.

RBF kernel function with parameters γ and C is defined as follows:

$$K\left(x_i, x_j\right) = \exp\left(-\gamma |x_i - x_j|^2\right) \tag{15}$$

2.4.2. K-nearest neighbor (KNN)

K-nearest neighbor (KNN) is an instance based classification method that has been effectively used in the area of machine learning and pattern recognition [77,78]. KNN is a non-parametric approach that does not frame any comprehensive model having no prior information about the training samples [79]. As an alternative, KNN finds the instances in the input feature space that is probably fit in a similar class. Therefore, the lazy learning nature of KNN makes it more efficient than eager learning for the unseen data samples [80]. Symptomatically, KNN classifies a data sample to the class, which seems most persistently among its nearest neighbor samples. It uses the Euclidian distance for measuring the distance among the instances of a feature space [81]. The calculated values of the neighbors are arranged in ascending order as $d_i < d_{i+1}$ where $i = 1$, 2, 3, …k, and k is the total number of instances in the feature vector.

2.4.3. Random forest (RF)

Random forest (RF) is an ensemble classification algorithm that was initially introduced by Breiman in 2001 [82]. RF has been effectively utilized in the area of machine learning for evaluating different classification and regression problem [83,84]. RF is a supervised learning procedure that is intrinsically capable of evaluating both binary class and multiclass problems [85]. RF uses a statistical Bootstrap technique, to construct multiple decision trees through a random selection of the data samples from training data [86]. Hence, a "forest" is grown having a large number of trees. The various number of predictors are used in order to find the best split at each node of the tree [87]. The random selection nature of RF made him efficient to reduce biases and minimize correlation among unpruned trees. Moreover; it also reduces variance by using ensemble pruned trees. Each prediction tree is a singleton, in the most frequently occurring class at input level [88]. Finally, the majority voting technique is used to generate an optimum output based on combing the predictions of each individual assumption. In this work, the number of trees is 100 and the number of iterations is 200.

2.5. Performance evaluation parameters

In machine learning, the effectiveness of an intelligent computational model is evaluated using different parameters [89,90]. Whereas, true and false predicted outcomes of the classification algorithm is keep in a confusion matrix [91]. Usually, in various prediction methods, accuracy is used to examine the strength of hypothesis learners, although the only accuracy is not sufficient to evaluate the efficiency of a prediction model [92,93]. However, a set of four metrics were introduced on the basis of Chou's symbols utilized for studying protein signal peptides and further these metrics were adopted by a series of publications [94–102]). However, the introduced metrics are only valid for single label systems; in case of multi-label systems (where a sample may instantaneously belong to various classes), such type of systems are more frequently existed in system biology [103–106], system medicine [107] and biomedicine [108], for which a complete different set of metrics are introduced [109]. In this model, the following performance measures are applied to accurately examine ACPs and non-ACPs.

$$Acc = 1 - \frac{ACP^{+}_{-} + ACP^{-}_{+}}{ACP^{+} + ACP^{-}} \tag{16}$$

$$Sen = 1 - \frac{ACP^{+}_{-}}{ACP^{+}} \tag{17}$$

$$Spe = 1 - \frac{ACP^{-}_{+}}{ACP^{-}} \tag{18}$$

$$MCC = \frac{1 - \left(\frac{ACP^{+}_{-} + ACP^{-}_{+}}{ACP^{+} + ACP^{-}}\right)}{\sqrt{\left(1 + \frac{ACP^{-}_{+} + ACP^{+}_{-}}{ACP^{+}}\right)\left(1 + \frac{ACP^{+}_{-} + ACP^{-}_{+}}{ACP^{-}}\right)}} \tag{19}$$

Table 1
Performance results of feature extraction techniques using Main dataset.

Methods	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
QSO	**SVM**	**96.52**	**76.01**	**98.58**	**0.78**
	KNN	92.96	80.44	94.22	0.65
	RF	96.28	63.55	99.55	0.75
CTF	SVM	95.68	61.33	99.11	0.71
	KNN	89.85	57.31	92	0.43
	RF	94.58	41.78	99.87	0.62
Geary Autocorrelation Descriptor	SVM	92.20	44.11	97.31	0.49
	KNN	91.39	40.52	96.03	0.41
	RF	92.88	44.88	98.28	0.46

In the above equations; ACP^{+} represents the number of anticancer peptides and ACP^{-} denotes non-anticancer peptides. Similarly, ACP^{+}_{-} are those anticancer peptides that are falsely classified as non-anticancer and ACP^{-}_{+} are non-anticancer peptides that are falsely predicted as anticancer peptides.

The graphical illustration of the proposed model is provided in Fig. 1. However, flowchart or graphically representation methods can provide an instinctive vision and also useful for analyzing the intricate correlation in the study of medical and biological systems, as mentioned by a series of essential biological topics, see, e.g., Refs. [110–113].

3. Results

In bioinformatics and machine learning, the predicted outcomes of a computational model can be evaluated using various statistical based cross-validation tests [113] such as sub-sampling test, independent test and jackknife and k-fold cross-validation test [114,115]. Among these tests; the results of the K-fold cross-validation test are fairly unbiased and possess minimum variance [116]. Therefore, in this study, the K-fold cross-validation was used to investigate the performance of the prediction model. Whereas, k-fold test divides the data ‘n’ into k-equal size (or nearly equal) folds. Subsequently, the training and testing of K iterations are performed, such that within each iteration different fold of data is held out for a testing purpose and the remaining ‘K-1’ number of folds are used training. In this section, we will briefly explain the performance outcomes of extracted feature spaces using different hypotheses learners.

3.1. Analysis of learning hypotheses using main dataset

The prediction rates of the proposed methods using the main dataset are listed in Table 1. Geary autocorrelation based feature vector obtained 16 features against each peptide sequence and obtained an accuracy of 92.88%, with the sensitivity of 44.86%, specificity of 98.28% and MCC of 0.47 using RF. In contrast, CTF based feature extraction gathered 343 efficient features against each peptide sequence. CTF feature space using SVM perform better than Geary autocorrelation and achieved the success rate of 95.68%, with sensitivity, specificity, and MCC of 61.33%, 99.11%, and 0.71, respectively. On the other hand, the effective features of QSO reported the highest accuracy of 96.52%, with the sensitivity of 76.01%, specificity of 98.58% and MCC of 0.78.

3.2. Analysis of main dataset after feature selection

The classification results of the main dataset after applying feature selection are illustrated in Table 2. In order to improve the classification performance, CTF feature space is minimized using PCA and finally, 300 efficient and irredundant features are selected. Likewise, the dimensionality of Geary autocorrelation is reduced to 12*2025. The reduced feature set using SVM obtained an accuracy of 92.48%, the sensitivity of 43.55%, and specificity of 96.93% and MCC of 0.47. On other hand, QSO feature space is also reduced to 35 consistent features and achieved the remarkable performance accuracy of 96.91% using SVM with sensitivity, specificity, and MCC of 77.32%, 98.12%, and 0.79, respectively.

Table 2
Performance results using Main dataset after feature selection.

Methods	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
QSO	**SVM**	**96.91**	**77.32**	**98.12**	**0.79**
	KNN	93.21	79.11	94.62	0.65
	RF	95.64	57.54	98.64	0.70
CTF	SVM	95.35	60.44	98.93	0.69
	KNN	90.18	54.34	94	0.44
	RF	93.17	35.22	95.56	0.48
Geary Autocorrelation Descriptor	SVM	92.48	43.55	96.93	0.47
	KNN	91.51	43.11	94.40	0.42
	RF	92.24	38.32	96.93	0.39

3.3. Analysis of learning hypotheses using alternate dataset

The prediction performance of alternate dataset using our proposed model is given in Table 3. In comparison with all the classification results, Geary feature vector reported the accuracy of 86.10%, with 43.22% sensitivity, 93.95% specificity, and 0.36 MCC. Similarly, CTF features in conjunction with SVM performed better than Geary autocorrelation and achieved an accuracy of 89.42%, sensitivity, specificity, and MCC of 57.11%, 94.35%, 0.52, respectively. Similarly, QSO feature space using alternated dataset also achieved the highest performance results, which are the accuracy of 89.54%, the sensitivity of 66.45%, specificity of 91.93% and MCC of 0.52.

3.4. Analysis of alternate dataset after feature selection

The success rates of the alternate dataset using the proposed model are given in Table 4. The reduced feature set of CTF shows a slight improvement in the classification results and obtained an accuracy of 89.54%. Similarly, Geary autocorrelation after applying PCA performs better using SVM and reported accuracy of 87.10%. Finally, QSO feature

Table 3
Performance results of feature extraction techniques using Alternate dataset.

Methods	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
QSO	**SVM**	**89.54**	**66.45**	**91.93**	**0.52**
	KNN	82.15	64.89	84.98	0.42
	RF	88.29	56.55	92.31	0.50
CTF	SVM	89.42	57.11	94.35	0.52
	KNN	82.78	53.79	87.54	0.37
	RF	89.31	41.88	96.03	0.48
Geary Autocorrelation Descriptor	SVM	86.10	43.22	93.95	0.36
	KNN	85.35	41.33	92.57	0.36
	RF	86.91	30.88	93.08	0.32

Table 4
Performance results using Alternate dataset after feature selection.

Methods	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
QSO	**SVM**	**88.35**	**52.33**	**92.75**	**0.48**
	KNN	81.78	64.44	84.62	0.41
	-RF	88.67	36.78	96.67	0.44
CTF	SVM	89.54	49.66	94.57	0.51
	KNN	82.34	56.89	86.52	0.38
	RF	86.29	24.19	91.13	0.27
Geary Autocorrelation Descriptor	SVM	87.10	41.55	95.23	0.38
	KNN	80.84	50.66	85.79	0.33
	RF	86.28	28.77	94.86	0.28

Table 5
Performance comparisons of our proposed model with existing methods.

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
This study [Main Dataset]	96.91	77.32	98.12	0.79
Tyagi et al.; [a]	92.65	74.67	94.44	0.61
This study [Alternate Dataset]	**89.54**	**66.45**	**91.93**	**0.52**
Tyagi et al.; [a]	76.38	65.22	78.08	0.33

[a] Tyagi, A., et al., *In silico models for designing and discovering novel anticancer peptides*. Scientific reports, 2013. 3: p. 2984.

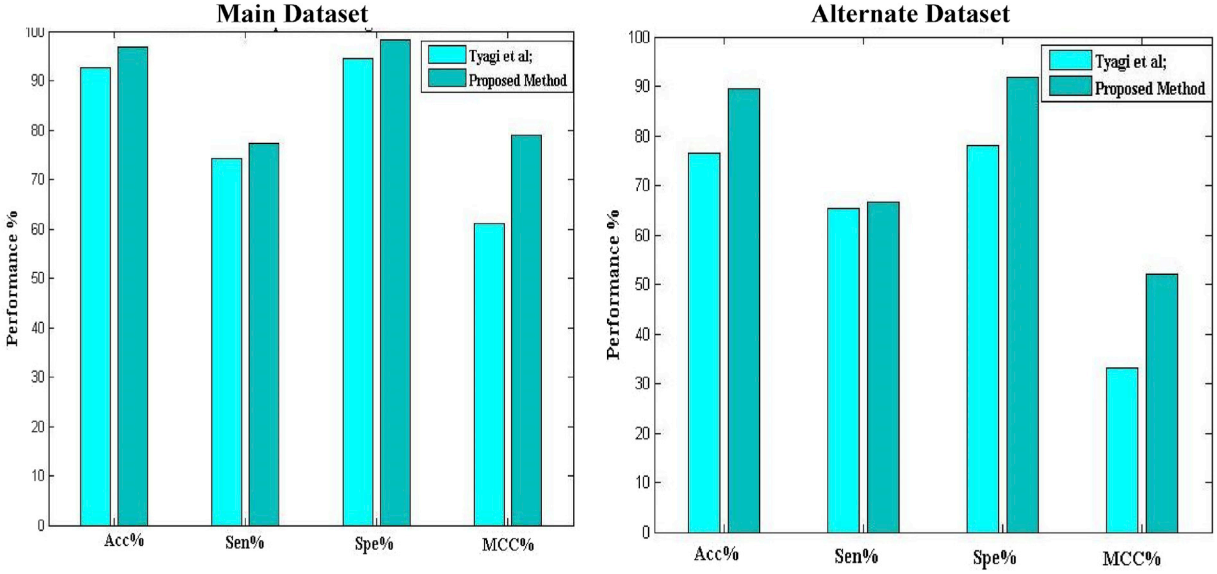

Fig. 2. Performance comparison of proposed and existing method using Main Dataset and Alternate Dataset.

space after feature selection does not show any significant improvement due to lack of sequence order information.

3.5. *Performance comparison of the proposed model with existing approaches*

The predicted outcomes of our proposed model in comparison with the existing state of art techniques are illustrated in Table 5. Initially; Tyagi et al.; used the similar datasets by applying SAAC and binary profile base model to identify ACPs [11]. After evaluating the experimental results it was found that Tyagi et al.; model using the main dataset achieved an accuracy of 92.65%, with the sensitivity of 74.67%, specificity of 94.44% and MCC of 0.61. In contrast, our proposed method reported the highest prediction rate of 96.91% with sensitivity, specificity, and MCC of 77.32%, 98.12% and 0.79, respectively. Similarly; in case of the alternate dataset, our proposed model performed efficiently and achieved 13% higher prediction rate than the existing model. Comparsion between proposed model and existing model on both datasets are illustrated in Fig. 2.

4. Discussion

Measuring the effectiveness of anticancer peptides over the traditional methods for cancer treatment, it is highly recommended to propose an automatic and intelligent model to accurately identify ACPs. Therefore, three diverse nature feature extraction schemes are used in this paper to extract nominal features from ACPs datasets. Among these methods, QSO performed outstanding due to its valuable and efficient sequence order features. QSO descriptors calculate chemical distance information and physiochemical properties using Grantham distance matrix and Schneider-Wrede distance matrix, respectively. In this work, QSO based features are collected from peptide sequences by keeping the lag value = 5 and weighting factor = 0.1. Moreover, the extracted feature space of QSO is then evaluated using several hypothesis learners among which, SVM achieved the remarkable performance results on all evaluation parameters in comparison with existing methods used in literature so far. On other hand, our proposed model also effectively deals with computational cost, by reducing the dimensionality of the extracted feature space using PCA. Thus, it is found that after applying PCA the performance results of both benchmark datasets does not affect.

5. Conclusions

In this paper, a reliable and effective intelligent model has been proposed for the accurate classification of ACPs. Three diverse nature feature representation methods namely: QSO, CTF and Geary Autocorrelation Descriptor are used to collect high discriminative information's from peptides sequences. Furthermore, PCA is applied to eliminate redundant features from extracted feature spaces. Finally, the predicted outcomes of the proposed model are measured using three different hypothesis learners such as; SVM, RF, and KNN. After investigating the predicted outcomes of the classification learners, it is observed that our proposed model achieved the highest performance results using both datasets than the existing models in the literature.

As pointed out in recent publications that [117], user-friendly and publicly accessible web-predictors provide the future direction for evaluating several essential computational analyses and innovation (see, e.g., Refs. [118–122].

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of competing interest

Authors Shahid Akbar, Ateeq ur Rehman and Maqsood Hayat declare that they have no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.103912.

References

[1] J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers, D.M. Parkin, Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008, Int. J. Cancer 127 (2010) 2893–2917.

[2] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, CA A Cancer J. Clin. 66 (2016) (2016) 7–30.

[3] P. Kanavos, The rising burden of cancer in the developing world, Ann. Oncol. 17 (2006) viii15–viii23.

[4] J. Thundimadathil, Cancer treatment using peptides: current therapies and future prospects, J. Amino Acids (2012) 2012.

[5] F. Harris, S.R. Dennison, J. Singh, D.A. Phoenix, On the selectivity and efficacy of defense peptides with respect to cancer cells, Med. Res. Rev. 33 (2013) 190–234.

[6] I. Fabregat, J. Fernando, J. Mainez, P. Sancho, TGF-beta signaling in cancer treatment, Curr. Pharmaceut. Des. 20 (2014) 2934–2947.

[7] S. Karbalaeemohammad, H. Naderi-Manesh, Two novel anticancer peptides from Aurein1. 2, Int. J. Pept. Res. Ther. 17 (2011) 159–164.

[8] F. Khan, S. Akbar, A. Basit, I. Khan, H. Akhlaq, Identification of anticancer peptides using optimal feature space of chou's split amino acid composition and support vector machine, in: Proceedings of the 2017 4th International Conference on Biomedical and Bioinformatics Engineering, ACM, 2017, pp. 91–96.

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

[10] B. Manavalan, S. Basith, T.H. Shin, S. Choi, M.O. Kim, G. Lee, MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 8 (2017) 77121.

[11] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, G. Raghava, In silico models for designing and discovering novel anticancer peptides, Sci. Rep. 3 (2013) 2984.

[12] F.-M. Li, X.-Q. Wang, Identifying anticancer peptides by using improved hybrid compositions, Sci. Rep. 6 (2016) 33910.

[13] S. Akbar, M. Hayat, M. Iqbal, M.A. Jan, iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space, Artif. Intell. Med. 79 (2017) 62–70.

[14] M. Kabir, M. Arif, S. Ahmad, Z. Ali, Z.N.K. Swati, D.-J. Yu, Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information, Chemometr. Intell. Lab. Syst. 182 (2018) 158–165.

[15] S. Vijayakumar, P. Lakshmi, ACPP: a web server for prediction and design of anti-cancer peptides, Int. J. Pept. Res. Ther. 21 (2015) 99–106.

[16] Z. Hajisharifi, M. Piryaiee, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou' s pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theor. Biol. 341 (2014) 34–40.

[17] L. Xu, G. Liang, L. Wang, C. Liao, A novel hybrid sequence-based model for identifying anticancer peptides, Genes 9 (2018) 158.

[18] C.T. Zhang, K.C. Chou, An optimization approach to predicting protein structural class from amino acid composition, Protein Sci. 1 (1992) 401–408.

[19] K.C. Chou, Y.D. Cai, Prediction and classification of protein subcellular location—sequence-order effect and pseudo amino acid composition, J. Cell. Biochem. 90 (2003) 1250–1260.

[20] K.-C. Chou, D.W. Elrod, Bioinformatical analysis of G-protein-coupled receptors, J. Proteome Res. 1 (2002) 429–433.

[21] L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. Cai, K.-C. Chou, Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties, PLoS One 6 (2011), e14556.

[22] Y.-D. Cai, K.-Y. Feng, W.-C. Lu, K.-C. Chou, Using LogitBoost classifier to predict protein structural classes, J. Theor. Biol. 238 (2006) 172–176.

[23] S. Akbar, M. Hayat, M. Kabir, M. Iqbal, iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins, Lett. Org. Chem. 16 (2019) 294–302.

[24] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins: Structure, Function, and Bioinformatics 43 (2001) 246–255.

[25] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′ s general PseAAC, J. Theor. Biol. 364 (2015) 284–294.

[26] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, Sci. Rep. 7 (2017) 42362.

[27] K.-C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337–2358.

[28] M. Novković, J. Simunić, V. Bojović, A. Tossi, D. Juretić, DADP: the database of anuran defense peptides, Bioinformatics 28 (2012) 1406–1407.

[29] G. Wang, X. Li, Z. Wang, APD2: the updated antimicrobial peptide database and its application in peptide design, Nucleic Acids Res. 37 (2008) D933–D937.

[30] S. Thomas, S. Karnik, R.S. Barai, V.K. Jayaraman, S. Idicula-Thomas, CAMP: a useful resource for research on antimicrobial peptides, Nucleic Acids Res. 38 (2009) D774–D780.

[31] UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2016) D158–D169.

[32] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.

[33] H.-B. Shen, K.-C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.

[34] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-Builder, A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions, Anal. Biochem. 425 (2012) 117–119.

[35] D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.

[36] P. Du, S. Gu, Y. Jiao, PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (2014) 3495–3506.

[37] K.-C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr. Proteomics 6 (2009) 262–274.

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

[39] M. Tahir, H. Tayara, K.T. Chong, iRNA-PseKNC (2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components, J. Theor. Biol. 465 (2019) 1–6.

[40] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68-e68.

[41] W. Chen, H. Lin, K.-C. Chou, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. Biosyst. 11 (2015) 2620–2634.

[42] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.-C. Chou, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.

[43] B. Liu, H. Wu, K.-C. Chou, Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nat. Sci. 9 (2017) 67.

[44] M. Hayat, A. Khan, Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, J. Theor. Biol. 271 (2011) 10–17.

[45] Z.U. Khan, M. Hayat, M.A. Khan, Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model, J. Theor. Biol. 365 (2015) 197–203.

[46] Z.-H. You, K.C. Chan, P. Hu, Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest, PLoS One 10 (2015), e0125811.

[47] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, X. Luo, Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding, BMC Bioinf. 17 (2016) 184.

[48] M. Khan, M. Hayat, S.A. Khan, N. Iqbal, Unb-DPC: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC, J. Theor. Biol. 415 (2017) 13–19.

[49] K.-C. Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem. Biophys. Res. Commun. 278 (2000) 477–483.

[50] M. Zhu, J. Dong, D. Cao, BioMedR: R/Bioconductor Package for Generating Various Molecular Representations for Chemicals, Proteins, DNAs/RNAs and their interactions, 2017.

[51] N. Xiao, Q. Xu, D. Cao, protr, Protein sequence feature extraction with R, R package version 0.2-0, URL, http://CRAN.R-project.org/package=protr, 2013.

[52] S.A. Ong, H.H. Lin, Y.Z. Chen, Z.R. Li, Z. Cao, Efficacy of different protein descriptors in predicting protein functional families, BMC Bioinf. 8 (2007) 300.

[53] B.A. van den Berg, M.J. Reinders, J.A. Roubos, D. de Ridder, SPiCE: a web-based tool for sequence-based protein classification and exploration, BMC Bioinf. 15 (2014) 93.

[54] H.D. Ismail, A. Jones, J.H. Kim, R.H. Newman, D.B. Kc, RF-Phos, A novel general Phosphorylation site prediction tool based on random Forest, BioMed Res. Int. (2016) 2016.

[55] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein–protein interactions based only on sequences information, Proc. Natl. Acad. Sci. 104 (2007) 4337–4341.

[56] X. Ma, J. Guo, X. Sun, Sequence-based prediction of RNA-binding proteins using random forest with minimum redundancy maximum relevance feature selection, BioMed Res. Int. (2015) 2015.

[57] J. Wang, L. Zhang, L. Jia, Y. Ren, G. Yu, Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences, Int. J. Mol. Sci. 18 (2017) 2373.

[58] S. Simeon, W. Shoombuatong, N. Anuwongcharoen, L. Preeyanon, V. Prachayasittikul, J.E. Wikberg, C. Nantasenamat, osFP: a web server for predicting the oligomeric states of fluorescent proteins, J. Cheminf. 8 (2016) 72.

[59] Y. Wang, Y. Tian, N. Deng, Distinguishing enzymes from non-enzymes via support vector machine, in: The Second International Symposium on Optimization and Systems Biology, Citeseer, 2008, pp. 166–173.

[60] Z. Yin, J. Tan, New encoding schemes for prediction of protein Phosphorylation sites, in: 2012 IEEE 6th International Conference on Systems Biology (ISB), IEEE, 2012, pp. 56–62.

[61] H. Wang, X. Hu, Accurate prediction of nuclear receptors with conjoint triad feature, BMC Bioinf. 16 (2015) 402.

[62] Y.-C. Wang, Y. Wang, Z.-X. Yang, N.-Y. Deng, Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context, BMC Syst. Biol. 5 (2011) S6.

[63] R.C. Geary, The contiguity ratio and statistical mapping, Inc. Statistician 5 (1954) 115–146.
[64] Y. Chen, New approaches for calculating Moran's index of spatial autocorrelation, PLoS One 8 (2013), e68336.
[65] Y. Liang, S. Liu, S. Zhang, Geary autocorrelation and DCCA coefficient: application to predict apoptosis protein subcellular localization via PSSM, Phys. A Stat. Mech. Appl. 467 (2017) 296–306.
[66] Z. Chen, C. Wu, Y. Zhang, Z. Huang, B. Ran, M. Zhong, N. Lyu, Feature selection with redundancy-complementariness dispersion, Knowl. Based Syst. 89 (2015) 203–217.
[67] I.M. Johnstone, D.M. Titterington, Statistical Challenges of High-Dimensional Data, The Royal Society Publishing, 2009.
[68] S. Khalid, T. Khalil, S. Nasreen, A survey of feature selection and feature extraction techniques in machine learning, in: 2014 Science and Information Conference, IEEE, 2014, pp. 372–378.
[69] D. Mladenić, Feature Selection for Dimensionality Reduction, International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection", Springer, 2005, pp. 84–102.
[70] S. Li, S. Oh, Improving feature selection performance using pairwise pre-evaluation, BMC Bioinf. 17 (2016) 312.
[71] K. Ahmad, M. Waris, M. Hayat, Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition, J. Membr. Biol. 249 (2016) 293–304.
[72] M.U. Ali, S. Ahmed, J. Ferzund, A. Mehmood, A. Rehman, Using PCA and Factor Analysis for Dimensionality Reduction of Bio-Informatics Data, 2017 arXiv preprint arXiv:1707.07189.
[73] X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: Tenth IEEE International Conference on Computer Vision, vol. 1, IEEE, 2005, pp. 1208–1213 (ICCV'05).
[74] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[75] S. Akbar, A. Ahmad, M. Hayat, Identification of fingerprint using discrete wavelet transform in conjunction with support vector machine, IJCSI 11 (2014), 1694-0814.
[76] S. Akbar, A. Ahmad, M. Hayat, F. Ali, Face recognition using hybrid feature space in conjunction with support vector machine, J. Appl. Environ. Biol. Sci 5 (2015) 28–36.
[77] D. Adeniyi, Z. Wei, Y. Yongquan, Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics 12 (2016) 90–108.
[78] R. Palaniappan, K. Sundaraj, S. Sundaraj, A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals, BMC Bioinf. 15 (2014) 223.
[79] J. Wu, A novel artificial neural network ensemble model based on K–Nearest neighbor nonparametric estimation of regression function and its application for rainfall forecasting, in: 2009 International Joint Conference on Computational Sciences and Optimization, IEEE, 2009, pp. 44–48.
[80] M. Tahir, 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, Mol. Biosyst. 12 (2016) 2587–2593.
[81] H. Liu, S. Zhang, J. Zhao, X. Zhao, Y. Mo, A new classification algorithm using mutual nearest neighbors, in: 2010 Ninth International Conference on Grid and Cloud Computing, IEEE, 2010, pp. 52–57.
[82] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[83] K. Fawagreh, M.M. Gaber, E. Elyan, Random forests: from early developments to recent advancements, Systems Science & Control Engineering: An Open Access Journal 2 (2014) 602–609.
[84] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (2002) 18–22.
[85] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.
[86] M. Waris, K. Ahmad, M. Kabir, M. Hayat, Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix, Neurocomputing 199 (2016) 154–162.
[87] M. Hayat, A. Khan, Mem-PHybrid: hybrid features-based prediction system for classifying membrane protein types, Anal. Biochem. 424 (2012) 35–44.
[88] M.F. Sabooh, N. Iqbal, M. Khan, M. Khan, H. Maqbool, Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC, J. Theor. Biol. 452 (2018) 1–9.
[89] A. Baratloo, M. Hosseini, A. Negida, G. El Ashal, Part 1: simple definition and calculation of accuracy, sensitivity and specificity, Emergency 3 (2015) 48–49.
[90] A.K. Dwivedi, Performance evaluation of different machine learning techniques for prediction of heart disease, Neural Comput. Appl. (2018) 1–9.
[91] Y. Jiao, P. Du, Performance measures in evaluating machine learning based bioinformatics predictors for classifications, Quantitative Biology 4 (2016) 320–330.
[92] S. Akbar, M. Hayat, iMethyl-STTNC, Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences, J. Theor. Biol. 455 (2018) 205–211.
[93] S.-H. Guo, E.-Z. Deng, L.-Q. Xu, H. Ding, H. Lin, W. Chen, K.-C. Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[94] K.-C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[95] K.C. Chou, Prediction of protein signal sequences and their cleavage sites, Proteins: Structure, Function, and Bioinformatics 42 (2001) 136–139.
[96] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ 1 (2013) e171.
[97] H. Lin, E.-Z. Deng, H. Ding, W. Chen, K.-C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.
[98] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.-C. Chou, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (2017) 4208.
[99] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, K.-C. Chou, iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, Genomics 111 (2019) 96–102.
[100] B. Liu, F. Yang, D.-S. Huang, K.-C. Chou, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, Bioinformatics 34 (2017) 33–40.
[101] B. Liu, F. Yang, K.-C. Chou, 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Mol. Ther. Nucleic Acids 7 (2017) 267–277.
[102] C.-J. Zhang, H. Tang, W.-C. Li, H. Lin, W. Chen, K.-C. Chou, iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition, Oncotarget 7 (2016) 69783.
[103] X. Xiao, X. Cheng, G. Chen, Q. Mao, K.-C. Chou, pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC, Genomics (2018).
[104] X. Cheng, X. Xiao, K.-C. Chou, pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, Genomics 110 (2018) 50–58.
[105] X. Xiao, X. Cheng, S. Su, Q. Mao, K.-C. Chou, pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins, Nat. Sci. 9 (2017) 330.
[106] X. Cheng, X. Xiao, K.-C. Chou, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315–321.
[107] X. Cheng, S.-G. Zhao, X. Xiao, K.-C. Chou, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (2017) 58494.
[108] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, K.-C. Chou, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.
[109] K.-C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.
[110] K.-C. Chou, S. Forsén, Graphical rules for enzyme-catalysed rate laws, Biochem. J. 187 (1980) 829–835.
[111] G.-P. Zhou, The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism, J. Theor. Biol. 284 (2011) 142–148.
[112] K.-C. Chou, W.-Z. Lin, X. Xiao, Wenxiang: a web-server for drawing wenxiang diagrams, Nat. Sci. 3 (2011) 862.
[113] K.-C. Chou, Graphic rule for drug metabolism systems, Curr. Drug Metabol. 11 (2010) 369–378.
[114] E. Saghapour, S. Kermani, M. Sehhati, A novel feature ranking method for prediction of cancer stages using proteomics data, PLoS One 12 (2017), e0184203.
[115] T.-T. Wong, N.-Y. Yang, Dependency analysis of accuracy estimates in k-fold cross validation, IEEE Trans. Knowl. Data Eng. 29 (2017) 2417–2427.
[116] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, J. Mach. Learn. Res. 5 (2004) 1089–1105.
[117] K.-C. Chou, H.-B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63.
[118] X. Cheng, W.-Z. Lin, X. Xiao, K.-C. Chou, pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC, Bioinformatics 35 (2018) 398–406.
[119] X. Xiao, X. Cheng, G. Chen, Q. Mao, K. Chou, pLoc_bal-mVirus: predict subcellular localization of multi-label virus proteins by PseAAC and IHTS treatment to balance training dataset, Med. Chem. (2018). Shariqah (United Arab Emirates.
[120] X. Cheng, X. Xiao, K.-C. Chou, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, Bioinformatics 34 (2017) 1448–1456.
[121] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.-C. Chou, iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther. Nucleic Acids 7 (2017) 155–163.
[122] X. Cheng, S.-G. Zhao, W.-Z. Lin, X. Xiao, K.-C. Chou, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (2017) 3524–3531.