International Journal of
Molecular Sciences

Article

mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides

Vinothini Boopathi [1], Sathiyamoorthy Subramaniyam [2,3], Adeel Malik [4], Gwang Lee [5,*], Balachandran Manavalan [5,*] and Deok-Chun Yang [1,*]

[1] Graduate School of Biotechnology, College of Life Science, Kyung Hee University, Yongin-si 17104, Gyeonggi-do, Korea; vinothini9327@gmail.com
[2] Research and Development Center, Insilicogen Inc., Yongin-si 16954, Gyeonggi-do, Korea; moorthy@insilicogen.com
[3] Department of Biotechnology, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu 641048, India
[4] Department of Microbiology and Molecular Biology, College of Bioscience and Biotechnology, Chungnam National University, Daejeon 34134, Korea; adeel@procarb.org
[5] Department of Physiology, Ajou University School of Medicine, Suwon 443380, Korea
* Correspondence: glee@ajou.ac.kr (G.L.); bala@ajou.ac.kr (B.M.); dcyang@khu.ac.kr (D.-C.Y.)

Received: 15 March 2019; Accepted: 18 April 2019; Published: 22 April 2019

Abstract: Anticancer peptides (ACPs) are promising therapeutic agents for targeting and killing cancer cells. The accurate prediction of ACPs from given peptide sequences remains as an open problem in the field of immunoinformatics. Recently, machine learning algorithms have emerged as a promising tool for helping experimental scientists predict ACPs. However, the performance of existing methods still needs to be improved. In this study, we present a novel approach for the accurate prediction of ACPs, which involves the following two steps: (i) We applied a two-step feature selection protocol on seven feature encodings that cover various aspects of sequence information (composition-based, physicochemical properties and profiles) and obtained their corresponding optimal feature-based models. The resultant predicted probabilities of ACPs were further utilized as feature vectors. (ii) The predicted probability feature vectors were in turn used as an input to support vector machine to develop the final prediction model called mACPpred. Cross-validation analysis showed that the proposed predictor performs significantly better than individual feature encodings. Furthermore, mACPpred significantly outperformed the existing methods compared in this study when objectively evaluated on an independent dataset.

Keywords: anticancer peptides; support vector machine; feature selection; optimal features; sequential forward search

1. Introduction

The complex process by which normal cells are transformed to abnormal cancer cells is known as carcinogenesis or tumorigenesis [1]. Such processes may be attributed to several factors, such as hereditation [2], environment [3], or a change in the physiological microenvironment of the affected cells [4]. Thus, most cancers—regardless of the driving factors—are distinguished by the gradual accumulation of genetic modifications in the founder cells [5]. Generally, the division and differentiation of normal cells are strictly regulated by several signaling pathways. However, sometimes normal cells escape these signals, leading to uncontrolled growth and proliferation, which ultimately leads to cancer [1]. According to the world health organization (WHO), the most common types of cancers are lung, liver, colorectal, stomach, prostate, skin and breast [https://www.who.int/news-room/fact-sheets/detail/cancer]. Due to cancer, millions of deaths have been reported each year from both developing and economically-advanced countries. In 2018, it was anticipated that about 18 million new cancer cases and over 9 million deaths could occur due to cancer [6], and these deaths could reach well over 13 million by 2030 [7]. In the United States (US) alone, approximately 1.7 million new cancer cases and over 600,000 cancer related deaths are estimated for 2019 [8].

Traditional methods for the treatment of cancer includes surgery, radiation therapy and chemotherapy, which may also depend on the location, stage of the disease, and the patient condition [9]. Despite advances, these methods are expensive and can often exhibit damaging effects on normal cells. Additionally, there is a growing concern that cancer cells may develop resistance to chemotherapy and molecularly-targeted therapies [10]. Moreover, cancer cells are known to develop multidrug resistance through a broad range of mechanisms, which not only makes these cells resistant to the drug in use for treatment, but also several other compounds [11]. As soon as the molecular mechanism behind cancer (or, as a matter of fact, any disease) is understood, the next logical step is to discover a desirable remedy for it [12]. Therefore, in view of the above, there is an urgent need to discover and design novel anti-cancer drugs to combat this deadly disease.

During the last few decades, the role of peptides as anti-cancer therapeutic agents has been promising, which is apparent from various strategies available to address the advancement of tumor growth and spreading of the disease [13]. These anti-cancer peptides (ACPs) have shown the potential to inactivate various types of cancer cells [11]. ACPs are short peptides (typically 5–50 amino acids in length) that exhibit high specificity, high tumor penetration and ease of synthesis and modification, in addition to low cost of production [14,15]. In general, most of the ACPs exhibit either an α-helical or a β-sheet conformation, however, in some cases, extended structures have also been identified [16]. ACPs can be classified into two major groups; i) peptides that are toxic to both cancerous and normal cells, exhibiting little evidence of selectivity, and ii) peptides that are toxic to cancer cells, but not to normal mammalian cells and erythrocytes [11]. The mechanisms by which these ACPs affect cancer cells are not yet completely understood. However, the role of membranolytic or non-membranolytic mechanisms are implicated [11]. Furthermore, the mechanisms that are involved in the inhibition of certain biological processes, such as angiogenesis, protein–protein interactions, signal transduction pathways, and gene expression, including the inhibition of enzymes or proteins, have also been highlighted [13].

Since most of the ACPs are derived from protein sequences [17], the discovery of novel ACPs for cancer treatment will be a focus of research for future studies. It is expected that the number of ACPs will increase with the rapid growth of protein sequences in public databases as a consequence of high-throughput sequencing projects [15]. Identification and development of novel ACPs from experimental methods is expensive and extremely time consuming. Therefore, it is essential to develop sequence-based computational methods to rapidly identify potential ACP candidates from the sequencing data prior to their synthesis. In this study, we constructed a lowest redundancy benchmark dataset and used this for the development of a prediction model. To develop a prediction model, we explored seven feature encodings, including amino acid composition (AAC), dipeptide composition (DPC), composition-transition-distribution (CTD), quasi-sequence-order (QSO), amino acid index (AAIF), binary profile (NC5), and conjoint triad (CTF). To exclude irrelevant features on each of the feature encodings, we applied a two-step feature selection protocol and identified their corresponding optimal feature-based models. Finally, the predicted probability obtained from seven feature encoding models were used as an input to a support vector machine (SVM) to construct the final model called mACPpred. Furthermore, our proposed method (mACPpred) achieved consistent performance on both benchmark and independent datasets.

2. Results

2.1. Performance of Various Feature Encodings

Firstly, we examined the capability of each feature encoding in classifying ACPs from non-ACPs. It should be noted that optimal machine-learning (ML) parameters for each feature encoding was obtained by conducting 10 independent 10-fold cross-validations. The best performance achieved by each feature encoding is shown in Figure 1. Results show that AAIF achieved the best performance with an accuracy of 88.72%, while AAC-, QSO-, DPC-, CTD-, CTF-, and NC5-based performance ranked at positions 2 to 7, respectively. Overall, seven feature encodings achieved a reasonable performance with an accuracy ranging between 81.0 and 88.7%. Furthermore, we observed that low-ranked feature encodings achieved the highest sensitivity and specificity. For instance, CTF achieved the highest sensitivity of 90.0%, which is 1.5–20% higher than the other encodings. Similarly, NC5 achieved the highest specificity of 91.35%, which is 1.06–18.0% higher than the other encodings. Although the basic nature of each feature encoding covers a different aspect of sequence information, each of these contribute towards better prediction. Therefore, it is essential to integrate these seven feature encoding-based models into a single model to overcome the limitations of each model and achieve a more balanced and stable performance.

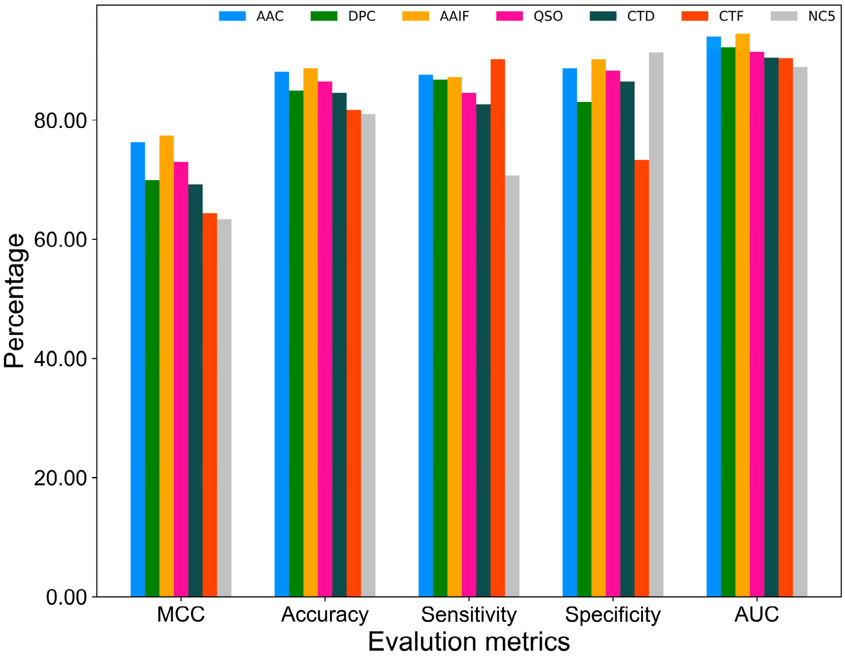

Figure 1. Performance of various feature encodings in a 10-fold cross-validation.

2.2. Comparison of SVM and Other Classifiers

To evaluate the effectiveness of SVM classifiers, we compared the performance of SVM-based classifiers against three other commonly used ML classifiers, namely random forest (RF), k-nearest neighbors (KNN) and logistic regression (LR), on seven feature encodings [18]. Using a 10-fold cross-validation test, the performance of the three other methods is shown in Table S1 and Figure 2. Results showed that SVM performed consistently better than the three other classifiers on six out of seven feature encodings. Precisely, the average accuracy achieved by SVM was ~1.1% higher than RF, ~2% higher than LR, and ~6% higher than KNN, indicating that SVM has a slight advantage over other methods in classifying ACPs from non-ACPs. Hence, we utilized only the SVM classifiers for further analysis.

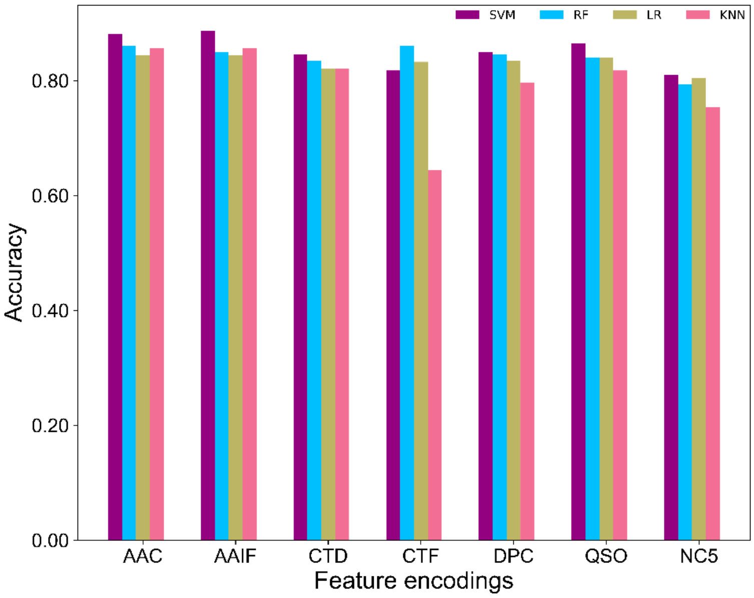

Figure 2. Comparison of SVM with other classifiers on seven different feature encodings.

2.3. *Selection of the Optimal Features for Each Encoding*

Since DPC, CTF and other encodings have a larger dimension, some of the features may be redundant or not all of them will be equally important. Therefore, it is mandatory to apply a feature selection protocol to remove redundant and irrelevant features. There are various feature selection techniques available in the literature [19–23], however, inspired by recent studies [24–26], we applied a two-step feature selection procedure to check whether it was able to reduce feature dimensions and improve performance. In particular, the F-score algorithm for ranking features (present in each feature encoding) was employed, followed by a sequential forward search to find the optimal feature set (Figure 3). Table 1 shows the number of features significantly reduced in the case of DPC (66.25%), CTF (58.82%), and NC5 (73%). On the other hand, a slight reduction can be observed in the case of AAIF (4.76%), QSO (1.0%), and CTD (10.62%). No reduction was witnessed in the case of AAC.

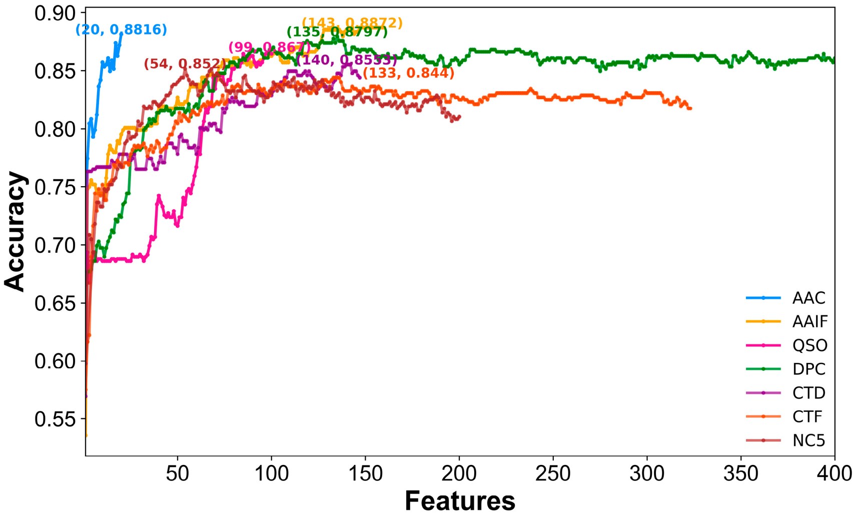

Figure 3. Sequential forward search for discriminating between anticancer peptides (ACPs) and non-ACPs. The maximum accuracy obtained from 10-fold cross-validation is shown for each feature encoding.

Table 1. The best performance achieved by various feature encodings using optimal features.

Feature Encoding	Dimension	MCC	Accuracy	Sensitivity	Specificity
AAC	20	0.763	0.882	0.876	0.887
DPC	135	0.762	0.880	0.838	0.921
CTD	140	0.711	0.853	0.842	0.865
AAIF	143	0.775	0.887	0.872	0.902
QSO	99	0.734	0.867	0.846	0.887
CTF	133	0.698	0.844	0.929	0.759
NC5	54	0.706	0.852	0.808	0.880

Next, we examined the performance of each feature encoding based on their optimal features and compared this with the respective control (using all features). Figure 4 shows a significant improvement in the performance for three feature encodings, NC5, DPC and CTF, by 4.18%, 3.08% and 2.63% respectively, as compared to their control. CTD and QSO improvement is marginal (<1%), while no improvement is observed in AAC and AAIF. Although no improvement was observed in AAIF, the number of feature dimensions is slightly reduced. In the case of AAC, all the features are equally important for achieving the best performance.

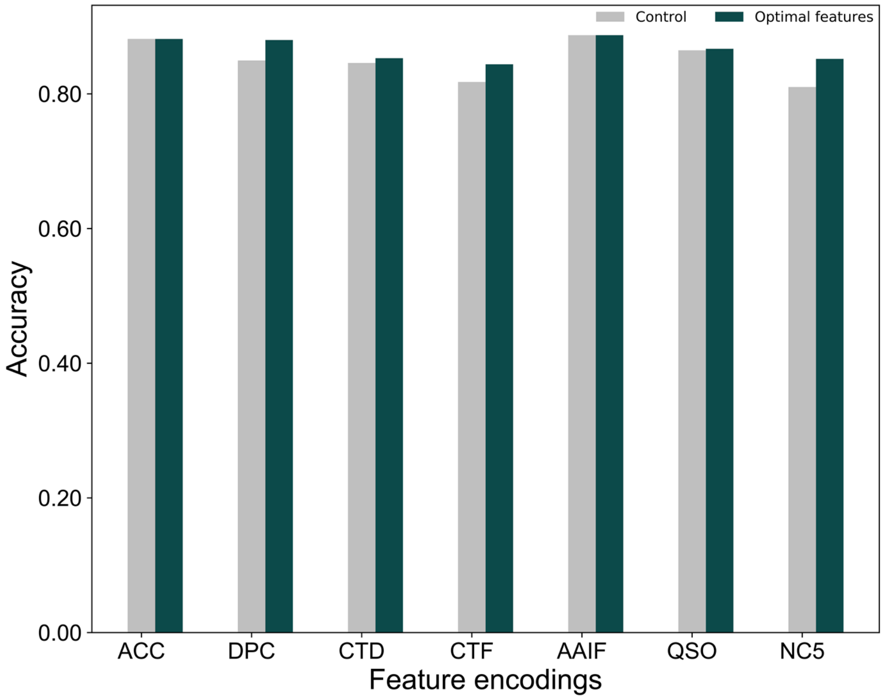

Figure 4. Performance comparison between the optimal feature set-based model against the respective controls (using all features).

To examine whether the optimal features are better than the excluded features for each feature encodings, we developed excluded features-based prediction models using the procedure as described in Section 2.1, and compared their performance with the control (using all features) and the optimal features. Notably, only four feature encodings (CTD, CTF, DPC, and NC5) were used for this analysis, while the remaining three feature encodings (AAC, AAIF, and QSO) were excluded considering the size of the optimal feature dimension and the control were similar. Figure S1 shows that the optimal feature-based models are consistently better than the control and also excluded feature-based models. Explicitly, the average accuracy achieved by the optimal feature-based models is 16.8% higher than the excluded feature-based models, and 2.7% higher than the control. This indicates that a two-step feature selection protocol selected more important features, thereby contributing to an improved performance. The optimal features for each feature encoding are provided in Table S2.

2.4. Construction of the Final Predictor

The optimal feature-based model obtained for each feature encoding was used in the development of a final prediction model. Some of the previous methods used hybrid features (a linear combination of various feature encodings) as an input to a ML classifier for the development of a prediction model without any feature selection techniques [27]. However, we considered only the predicted probability of ACPs (values in the range of 0.0 to 1.0) from seven individual optimal models as input features to SVM and developed a final prediction model called mACPpred. Our proposed predictor achieves an Matthews correlation coefficient (MCC), accuracy, sensitivity, specificity and area under the curve (AUC) of 0.836, 0.917, 0.891, 0.944, and 0.968, respectively. To show the effectiveness of mACPpred, we compared its performance with seven feature encoding predictors (Figure 5A). Specifically, the MCC and accuracy of the proposed predictor was 4.6–13.8% and 3.5–7.3% higher than the individual predictors, thus indicating the effectiveness of our approach by integrating various feature encodings, which in turn contributes to an improved performance.

It might be possible that methods employing hybrid features (a combination of different feature encodings) perform better than the current approach because they utilize multiple elements and also complete the feature space. To investigate this possibility, we developed six hybrid-feature-based models using the following procedure: (i) Seven feature encodings were ranked according to the accuracy obtained from base-line models (Figure 1) and incorporated with AAIF one by one (H1: AAIF+AAC; H2: AAIF+AAC+QSO; H3: AAIF+AAC+QSO+DPC; H4: AAIF+AAC+QSO+DPC+CTD; H5: AAIF+AAC+QSO+DPC+CTD+CTF; H6: AAIF+AAC+QSO+DPC+CTD+CTF+NC5). Each of the hybrid features were used as an input to SVM and their corresponding models were developed using the same procedure as described in Section 2.1. Figure 5B shows the performance comparison of mACPpred with the hybrid-feature-based models, where mACPpred performed better with an MCC and accuracy value 3.46–9.5% and 1.7–4.7% higher than the hybrid models, respectively, thereby demonstrating the advantage of our approach in achieving the best performance.

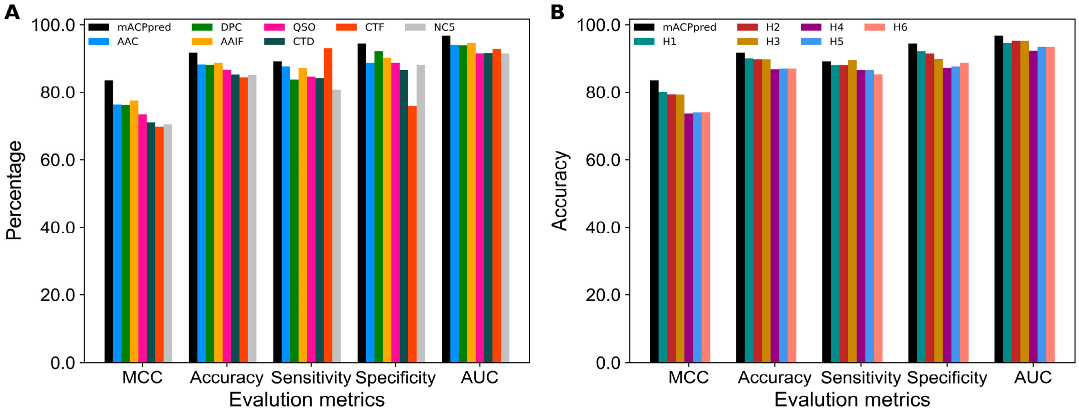

Figure 5. (**A**) Performance comparison of mACPpred with the single feature models, based on optimal features. (**B**) Performance comparison between mACPpred and hybrid features-based models.

2.5. Performance Comparison on the Independent Dataset

There are several examples where the prediction model showed an excellent performance during cross-validation. However, these performances are not transferrable while evaluating an independent dataset. Hence, an independent evaluation is needed to validate the robustness of the proposed method. Importantly, the independent dataset constructed in this study did not share greater than 90% sequence identity with our training dataset and other existing methods' training datasets. We compared the performances of mACPpred with the previous methods, such as MLACP and iACP. It should be noted that MLACP contains two prediction models based on RF (RFACP) and SVM (SVMACP), and both the models were considered for comparison.

Table 2 shows that mACPpred achieves an MCC, accuracy, sensitivity, specificity, and AUC of 0.829, 0.914, 0.885, 0.943, and 0.967, respectively. More specifically, the MCC and accuracy of mACPpred is 23.7–49.1% and 14.6–33.4% higher, respectively, than the other methods compared in this study, demonstrating that the proposed method is capable of achieving an encouraging performance. Note that it is hard to get statistical estimation from the above threshold-based comparison. Hence, we utilized rank-based comparison using ROC [28], where two AUC values of different methods were assessed by a two-tailed test, from which the p value for the observed differences were obtained [29]. Table 2 and Figure 6 shows that the mACPpred significantly outperformed the existing predictors on the independent dataset.

Table 2. Performance of various methods on the independent dataset.

Methods	MCC	Accuracy	Sensitivity	Specificity	AUC	p Value
mACPpred	0.829	0.914	0.885	0.943	0.967	–
SVMACP [27]	0.592	0.768	0.554	0.981	0.896	0.000382
RFACP [27]	0.511	0.707	0.414	1.000	0.891	0.000401
iACP [30]	0.338	0.667	0.580	0.753	0.747	<0.00001

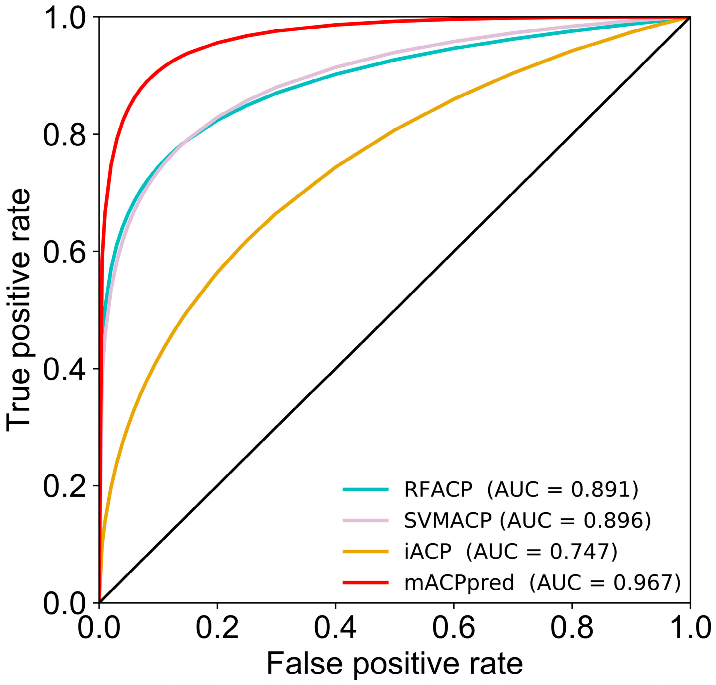

Figure 6. Comparison of binormal receiver operating characteristics (ROC) curves for ACPs prediction using different methods on independent dataset.

2.6. Webserver Implementation

mACPpred webserver is freely accessible at the following link: www.thegleelab.org/mACPpred. Users can upload or paste query peptide sequences in the FASTA format, and after submitting peptide sequences, retrieve results in a separate interface. All datasets used in this study can be downloaded from the following link: http://thegleelab.org/mACPpred/ACPData.html, to check the reproducibility of our findings.

3. Discussion

In this study, we developed a novel predictor called mACPpred to predict ACPs from the given peptide sequence. To develop a predictor, a two-step feature selection protocol was applied on seven feature encodings (AAC, DPC, CTD, CTF, AAIF, QSO, and NC5) to obtain optimal feature-based prediction models, whose predicted probabilities of ACPs were further used as a feature vector. Finally, the probabilistic feature vector was used as an input to SVM for the development of a final prediction model. The benchmark and independent validation demonstrated that the mACPpred was able to clearly outperform existing predictors compared in this study for ACPs prediction. The novelty of our method is as follows: (i) The benchmark or training dataset has the lowest redundancy among the datasets reported in the literature; (ii) among various feature encodings employed in this study, this is the first instance where CTF and QSO are employed in ACP prediction, and (iii) most of the existing predictors either utilize single feature encodings or a combination of multiple feature encodings, hence, their feature dimension is very high. However, we have used only seven probabilistic features that cover a wide range of features (position specific, physicochemical, and compositional information). Basically, it transforms the complex high-dimensional feature into a low-dimensional one, further facilitating better discrimination between ACPs and non-ACPs.

Moreover, our approach can be applied to other sequence-based prediction problems, including post-translational modifications, peptide function predictions, and DNA/RNA function predictions. Although the proposed predictor has shown an excellent performance over the other methods, there is still room for improvement. This includes exploration of other ML algorithms such as decision tree-based [31,32] and neural network-based algorithms [33–35] on the same dataset, incorporation of novel features and computational approach as implemented in References [36–39], and increasing the size of the training dataset based on the future experimental data. Furthermore, we implemented our proposed algorithm in the form of user-friendly web-server (http://thegleelab.org/mACPpred) for the wider research community to use. We expect that mACPpred will be helpful for identifying novel potential ACPs.

4. Materials and Methods

4.1. Dataset Collection and Processing

We generated positive samples by utilizing the previously reported datasets of Tyagi et al. [40], Wei et al. [15], and Chen et al. [30], which contain 225, 250, and 288 (both training and independent datasets) experimentally verified ACPs, respectively. From this, we excluded sequences >50 amino acid residues because these may form outliers during prediction model development as very few peptides have larger than 50 amino acids. Subsequently, we applied a CD-HIT [41] threshold of 0.8 and excluded the redundant sequences, which resulted 266 positive samples. It should be noted that lower thresholds of the sequence identity (less than 50%) might reduce the sequence homology bias and could improve the model credibility. However, using higher threshold was necessary due to the smaller dataset size. For the collection of negative samples, we utilized 2250 samples reported by Tyagi et al. [40], wherein AMPs had been extracted from several databases including, APD, CAMP, DADP, for which no anticancer activity has been reported in the literature. Subsequently, we applied CD-HIT of 0.8 against the positive samples and among negative samples that resulted 2069 non-ACPs.

Training dataset: We generated a high-quality training dataset by selecting 266 ACPs and randomly selecting 266 non-ACPs from the 2069 non-ACPs set mentioned above. Generally, ML classifiers tend to produce unbiased performance on a balanced dataset [15], hence we selected an equal number of non-ACPs with ACPs. To the best of our knowledge, our training dataset has the lowest redundancy among the reported datasets, which was utilized to develop a prediction model.

Independent dataset: we manually collected positive samples (ACPs) from the following databases: DADP [42], DBAASP [43], DRAMP [44], and LAMP [45]. Subsequently, we applied a CD-HIT threshold of 0.9 among the collected peptides and also against the training dataset, which resulted in 157 ACPs.

Furthermore, 157 non-ACPs were randomly selected from Tyagi's negative dataset, mentioned above, which did not overlap with the training and independent positive datasets. This dataset was used to evaluate our prediction model. All the datasets utilized in this study can be found in the Supplementary Information.

4.2. Feature Extraction

To develop or train a prediction model, it is essential to formulate a diverse length of peptides as a fixed length of feature vectors. In this study, we explored seven different feature encodings that can be grouped into sequence-based features and physicochemical properties-based features, as described below:

4.2.1. Sequence-Based Features

The differences between peptides can be reflected by amino acid sequences, which includes composition, profiles, physicochemical properties, permutation and combination modes of amino acids. Hence, we extracted five types of sequence-based features: AAC; DPC, QSO, CTF, and NC5.

1. AAC has been widely used in numerous sequence-based prediction tasks [46], which represents the occurrence frequencies of 20 standard amino acids in a given peptide sequence that generates a 20-dimensional vector.
2. DPC encoding of the given peptide sequence results in a fixed length of a 400-dimensional feature vector that summarizes amino acids fraction, the sequence-order, and fragment information.
3. QSO encoding of the given peptide sequence results in a fixed length of a 100-dimensional feature vector, by measuring the physicochemical distance between the amino acids within the sequence. A detailed description of QSO feature encoding along with a set of equations has been provided in previous studies [47–49].
4. CTF encoding generates a 343-dimensional feature vector for a given peptide sequence by clustering amino acids into seven classes according to their dipoles and side-chains volumes. A detailed description of CTF with a set of equation has been provided in previous studies [50–52].
5. In NC5, each amino acid is encoded as a 20-dimensional 0/1 vector. For example, the amino acid of type A and type C are encoded as (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) and (0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), respectively. BPF(w) dimension is 20 × w, where w is sequence window length. Since the minimal length of the peptide in our dataset is 5, we fixed at this value and considered both N-and C- terminals (NC5) that generate a 200-dimensional vector.

4.2.2. Physicochemical Properties-Based Features

Physicochemical properties have been widely used and successfully applied in numerous prediction problems including proteins [23,24,36], RNA [53,54], and DNA [38,55,56]. Here, we used the following two types: AAIF and CTD; both represent a global composition of amino acid properties in a different perspective.

1. In AAIF, we used only eight high-quality amino acid indices as reported in a previous study [57], which are LIFS790101 [58], CEDJ970104 [59], MIYS990104 [60], NAKH920108 [61], TSAJ990101 [62], MAXF760101 [63], BIOV880101 [64], BLAM930101 [65]. AAIF generates a 160 (=20 amino acids * 8 properties) dimensional vector, which has been successfully applied in numerous sequence-based prediction tasks [66–68].
2. We used seven different types of physicochemical properties listed in Table 3 where 20 standard amino acids are classified into 3 different classes according to their attributes. In CTD, composition, transition, and distribution are respectively encoded as a 21, 21, 105-dimensional feature vector. A detailed description of CTD with a set of equations has been provided in previous studies [69,70].

Table 3. Classification of 20 amino acids according to the seven specific types of physicochemical properties.

Properties	**Class1**	**Class2**	**Class3**
Hydrophobicity	Polar E, D, K, N, Q, R	Neutral A, G, H, P, S, T, Y	Hydrophobicity C, L, V, I, M, F, W
Normalized Van der Waals volume	0–2.78 A, C, D, G, P, S, T	2.95–4.0 E, I, L, N, V, Q	4.03–8.08 M, H, K, F, R, Y, W
Polarity	4.9–6.2 L, I, F, W, C, M, V, Y	8.0–9.2 A, G, P, S, T	10.4–13.0 H, Q, R, K, N, E, D
Polarizability	0–0.108 A, D, G, S, T	0.128–0.186 C, E, I, L, P, Q, V, N	0.219–0.409 K, M, H, F, R, Y, W
Charge	Positive K, R	Neutral A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	Negative D, E
Secondary Structure	Helix A, E, H, K, L, M, Q, R	Strand V, I, Y, C, W, F, T	Coil D, G, N, P, S
Solvent Accessibility	Buried A, C, F, G, I, L, V, W	Exposed D, E, K, N, Q, R	Intermediate M, S, P, T, H, Y

4.3. Support Vector Machine

SVM is one of the powerful machine learning algorithms, which has been widely used in numerous fields within bioinformatics [21,22,53,71–73]. The *scikit-learn* (v.0.19.1) library in Python was used to implement SVM algorithm [74]. The main objective of SVM is to find the optimal hyperplane that can maximize the distance between two categories (positive and negative) in high-dimensional feature space. We used the radial basis function (RBF) as the kernel function of SVM because the preliminary analysis showed that RBF performed superior in comparison with the other three kernel functions (linear, polynomial, and sigmoid). Here, the grid search approach was employed to optimize the two parameters of RBF-SVM: The penalty parameter C and the kernel parameter γ. These parameters were tuned with the following search space:

$$\begin{cases} 2^{-5} \le C \le 2^{15} withstep \Delta C = 2 \\ 2^{-15} \le \gamma \le 2^{15} withstep \Delta\gamma = 2^{-1} \end{cases} \tag{1}$$

4.4. Ten-Fold Cross-Validation

Generally, three cross-validation (CV) methods, namely an independent data set test, a sub-sampling (or κ-fold CV) test, and a leave-one-out CV (LOOCV) test, are used to evaluate the anticipated success rate of a predictor [32,75]. In this study, we used a 10-fold CV to examine the proposed models. In the 10-fold CV, the benchmarking dataset was randomly partitioned into 10 subsets. One subset was used as a test set and the remaining nine subsets were used as the training sets. This procedure was repeated 10 times, with each subset being used once as a test set. The performance of the 10 corresponding results are averaged, with the outcome implying the performance of the classifier.

4.5. Feature Selection

Feature selection is one of the most important steps in developing ML-based models, and improve classification performance. In this study, we used the F-score algorithm along with a sequential forward search strategy to identify the optimal features [76]. Initially, the F-score algorithm was used to rank all the features and sort them from the highest to the lowest scores and thereby generate a ranked feature list. Later, features were added one by one from the ranked list, developing their corresponding predicting models. Finally, the feature subset that achieved the highest accuracy was regarded as the optimal features.

4.6. Performance Evaluation of ACPs Prediction

To evaluate the performance of the proposed method, the following four commonly used metrics [39,77–82] were employed: Sensitivity (SN), specificity (SP), accuracy (ACC), and the Matthews correlation coefficient (MCC), which are computed using the following formula:

$$\left\{\begin{array}{r} SN = \frac{TP}{TP+FN} \\ SP = \frac{TN}{TN+FP} \\ ACC = \frac{TP+FN}{TP+TN+FN+FP} \\ MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \end{array}\right. \tag{2}$$

where TP and TN are the number of ACPs correctly predicted and non-ACPs correctly predicted, respectively. FN is the number of ACPs predicted as non-ACPs, whereas FP is the number of non-ACPs predicted as ACPs.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/8/1964/s1.

Author Contributions: B.M., G.L. and D.-C.Y. conceived and designed the experiments; V.B., A.M., and S.S. performed the experiments; V.B., A.M., B.M. and D.-C.Y. wrote the paper.

Funding: This research was supported by the grant from Korea Institute of Planning & Evaluation for Technology in Food, Agriculture, Forestry & Fisheries (KIPET NO: 317007-3) and the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science, and Technology (2018R1D1A1B07049572).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Salehi, B.; Zucca, P.; Sharifi-Rad, M.; Pezzani, R.; Rajabi, S.; Setzer, W.N.; Varoni, E.M.; Iriti, M.; Kobarfard, F.; Sharifi-Rad, J. Phytotherapeutics in cancer invasion and metastasis. *Phytother. Res.* **2018**, *32*, 1425–1449. [CrossRef]
2. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **2014**, *505*, 302–308. [CrossRef]
3. Wild, C.P.; Scalbert, A.; Herceg, Z. Measuring the exposome: A powerful basis for evaluating environmental exposures and cancer risk. *Environ. Mol. Mutagen* **2013**, *54*, 480–499. [CrossRef] [PubMed]
4. Gillies, R.J.; Gatenby, R.A. Metabolism and its sequelae in cancer evolution and therapy. *Cancer J.* **2015**, *21*, 88–96. [CrossRef] [PubMed]
5. Storey, K.; Ryser, M.D.; Leder, K.; Foo, J. Spatial Measures of Genetic Heterogeneity During Carcinogenesis. *Bull. Math. Biol.* **2017**, *79*, 237–276. [CrossRef]
6. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef]
7. Boyle, P.; Levin, B. *World Cancer Report 2008*; IARC Press, International Agency for Research on Cancer: Lyon, France, 2008.
8. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.* **2019**, *69*, 7–34. [CrossRef] [PubMed]
9. Kakde, D.; Jain, D.; Shrivastava, V.; Kakde, R.; Patil, A. Cancer therapeutics-opportunities, challenges and advances in drug delivery. *J. Appl. Pharm. Sci.* **2011**, *1*, 1–10.
10. Holohan, C.; Van Schaeybroeck, S.; Longley, D.B.; Johnston, P.G. Cancer drug resistance: An evolving paradigm. *Nat. Rev. Cancer* **2013**, *13*, 714. [CrossRef]
11. Harris, F.; Dennison, S.R.; Singh, J.; Phoenix, D.A. On the selectivity and efficacy of defense peptides with respect to cancer cells. *Med. Res. Rev.* **2013**, *33*, 190–234. [CrossRef] [PubMed]
12. Malik, A.; Singh, H.; Andrabi, M.; Husain, S.A.; Ahmad, S. Databases and QSAR for cancer research. *Cancer Inform.* **2006**, *2*, 99–111. [CrossRef]
13. Thundimadathil, J. Cancer treatment using peptides: Current therapies and future prospects. *J. Amino Acids* **2012**, *2012*, 967347. [CrossRef]

14. Otvos, L., Jr. Peptide-based drug design: Here and now. *Methods Mol. Biol.* **2008**, *494*, 1–8. [PubMed]
15. Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: A sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34*, 4007–4016. [CrossRef]
16. Gaspar, D.; Veiga, A.S.; Castanho, M.A. From antimicrobial to anticancer peptides. A review. *Front. Microbiol.* **2013**, *4*, 294. [CrossRef]
17. Tyagi, A.; Tuknait, A.; Anand, P.; Gupta, S.; Sharma, M.; Mathur, D.; Joshi, A.; Singh, S.; Gautam, A.; Raghava, G.P. CancerPPD: A database of anticancer peptides and proteins. *Nucleic Acids Res.* **2015**, *43*, D837–D843. [CrossRef]
18. Stephenson, N.; Shane, E.; Chase, J.; Rowland, J.; Ries, D.; Justice, N.; Zhang, J.; Chan, L.; Cao, R. Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.* **2018**. [CrossRef]
19. Tan, J.-X.; Dao, F.-Y.; Lv, H.; Feng, P.-M.; Ding, H. Identifying Phage Virion Proteins by Using Two-Step Feature Selection Methods. *Molecules* **2018**, *23*, 2000. [CrossRef]
20. Cascio, D.; Taormina, V.; Raso, G. An Automatic HEp-2 Specimen Analysis System Based on an Active Contours Model and an SVM Classification. *Appl. Sci.* **2019**, *9*, 307. [CrossRef]
21. Manavalan, B.; Lee, J. SVMQA: Support-vector-machine-based protein single-model quality assessment. *Bioinformatics* **2017**, *33*, 2496–2503. [CrossRef] [PubMed]
22. Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Front. Microbiol.* **2018**, *9*, 476. [CrossRef] [PubMed]
23. Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: Support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* **2018**, *9*, 1944–1956. [CrossRef] [PubMed]
24. Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 412–420. [CrossRef]
25. Qiang, X.; Chen, H.; Ye, X.; Su, R.; Wei, L. M6AMRFS: Robust Prediction of N6-Methyladenosine Sites With Sequence-Based Features in Multiple Species. *Front. Genet.* **2018**, *9*, 495. [CrossRef] [PubMed]
26. Zhang, M.; Li, F.; Marquez-Lago, T.T.; Leier, A.; Fan, C.; Kwoh, C.K.; Chou, K.C.; Song, J.; Jia, C. MULTiPly: A novel multi-layer predictor for discovering general and specific types of promoters. *Bioinformatics* **2019**. [CrossRef]
27. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121–77136. [CrossRef] [PubMed]
28. Gabere, M.N.; Noble, W.S. Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics* **2017**, *33*, 1921–1929. [CrossRef] [PubMed]
29. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
30. Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895–16909. [CrossRef]
31. Manavalan, B.; Lee, J.; Lee, J. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS ONE* **2014**, *9*, e106542. [CrossRef]
32. Su, R.; Liu, X.; Wei, L.; Zou, Q. Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods* **2019**. [CrossRef] [PubMed]
33. Tang, H.; Cao, R.-Z.; Wang, W.; Liu, T.-S.; Wang, L.-M.; He, C.-M. A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomath.* **2017**, *10*, 1750050. [CrossRef]
34. Conover, M.; Staples, M.; Si, D.; Sun, M.; Cao, R. AngularQA: Protein Model Quality Assessment with LSTM Networks. *bioRxiv* **2019**, *560995*. [CrossRef]
35. Hou, J.; Wu, T.; Cao, R.; Cheng, J. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *bioRxiv* **2019**, *552422*. [CrossRef] [PubMed]
36. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G.; Hancock, J. mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* **2018**. [CrossRef] [PubMed]
37. Qiang, X.; Zhou, C.; Ye, X.; Du, P.F.; Su, R.; Wei, L. CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief Bioinform.* **2018**. [CrossRef]

38. Wei, L.; Luan, S.; Nagai, L.A.E.; Su, R.; Zou, Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **2018**. [CrossRef]
39. Cao, R.; Adhikari, B.; Bhattacharya, D.; Sun, M.; Hou, J.; Cheng, J. QAcon: Single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* **2017**, *33*, 586–588. [CrossRef] [PubMed]
40. Tyagi, A.; Kapoor, P.; Kumar, R.; Chaudhary, K.; Gautam, A.; Raghava, G.P. In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* **2013**, *3*, 2984. [CrossRef]
41. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680–682. [CrossRef]
42. Novkovic, M.; Simunic, J.; Bojovic, V.; Tossi, A.; Juretic, D. DADP: The database of anuran defense peptides. *Bioinformatics* **2012**, *28*, 1406–1407. [CrossRef]
43. Pirtskhalava, M.; Gabrielian, A.; Cruz, P.; Griggs, H.L.; Squires, R.B.; Hurt, D.E.; Grigolava, M.; Chubinidze, M.; Gogoladze, G.; Vishnepolsky, B.; et al. DBAASP v.2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.* **2016**, *44*, D1104–D1112. [CrossRef]
44. Fan, L.; Sun, J.; Zhou, M.; Zhou, J.; Lao, X.; Zheng, H.; Xu, H. DRAMP: A comprehensive data repository of antimicrobial peptides. *Sci. Rep.* **2016**, *6*, 24482. [CrossRef]
45. Zhao, X.; Wu, H.; Lu, H.; Li, G.; Huang, Q. LAMP: A Database Linking Antimicrobial Peptides. *PLoS ONE* **2013**, *8*, e66557. [CrossRef]
46. Usmani, S.S.; Kumar, R.; Bhalla, S.; Kumar, V.; Raghava, G.P.S. In Silico Tools and Databases for Designing Peptide-Based Vaccine and Drugs. *Adv. Protein Chem. Struct. Biol.* **2018**, *112*, 221–263.
47. Chou, K.C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.* **2000**, *278*, 477–483. [CrossRef]
48. Wang, J.; Li, J.; Yang, B.; Xie, R.; Marquez-Lago, T.T.; Leier, A.; Hayashida, M.; Akutsu, T.; Zhang, Y.; Chou, K.C.; et al. Bastion3: A two-layer ensemble predictor of type III secreted effectors. *Bioinformatics* **2018**. [CrossRef] [PubMed]
49. Wang, J.; Yang, B.; Leier, A.; Marquez-Lago, T.T.; Hayashida, M.; Rocker, A.; Zhang, Y.; Akutsu, T.; Chou, K.C.; Strugnell, R.A.; et al. Bastion6: A bioinformatics approach for accurate prediction of type VI secreted effectors. *Bioinformatics* **2018**, *34*, 2546–2555. [CrossRef]
50. Lin, T.W.; Wu, J.W.; Chang, D.T. Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins. *PLoS ONE* **2013**, *8*, e75940. [CrossRef]
51. Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4337–4341. [CrossRef]
52. Wang, J.; Zhang, L.; Jia, L.; Ren, Y.; Yu, G. Protein-Protein Interactions Prediction Using a Novel Local Conjoint Triad Descriptor of Amino Acid Sequences. *Int. J. Mol. Sci.* **2017**, *18*, 2373. [CrossRef] [PubMed]
53. Wei, L.; Chen, H.; Su, R. M6APred-EL: A Sequence-Based Predictor for Identifying N6-methyladenosine Sites Using Ensemble Learning. *Mol. Ther. Nucleic Acids* **2018**, *12*, 635–644. [CrossRef]
54. Zou, Q., Sr.; Xing, P.; Wei, L.; Liu, B. Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA. *RNA* **2018**, *25*, 205–218. [CrossRef]
55. Chen, W.; Lv, H.; Nie, F.; Lin, H. i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* **2019**. [CrossRef] [PubMed]
56. Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: Identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **2017**, *33*, 3518–3523. [CrossRef] [PubMed]
57. Saha, I.; Maulik, U.; Bandyopadhyay, S.; Plewczynski, D. Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids* **2012**, *43*, 583–594. [CrossRef]
58. Lifson, S.; Sander, C. Antiparallel and parallel beta-strands differ in amino acid residue preferences. *Nature* **1979**, *282*, 109–111. [CrossRef]
59. Cedano, J.; Aloy, P.; Perez-Pons, J.A.; Querol, E. Relation between amino acid composition and cellular location of proteins1. *J. Mol. Biol.* **1997**, *266*, 594–600. [CrossRef]
60. Miyazawa, S.; Jernigan, R.L. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* **1999**, *34*, 49–68. [CrossRef]
61. Sipos, L.; von Heijne, G. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* **1993**, *213*, 1333–1340. [CrossRef] [PubMed]

62. Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. The packing density in proteins: Standard radii and volumes. *J. Mol. Biol.* **1999**, *290*, 253–266. [CrossRef]
63. Maxfield, F.R.; Scheraga, H.A. Status of empirical methods for the prediction of protein backbone topography. *Biochemistry* **1976**, *15*, 5138–5153. [CrossRef]
64. Biou, V.; Gibrat, J.F.; Levin, J.M.; Robson, B.; Garnier, J. Secondary structure prediction: Combination of three different methods. *Protein Eng.* **1988**, *2*, 185–191. [CrossRef] [PubMed]
65. Blaber, M.; Zhang, X.J.; Matthews, B.W. Structural basis of amino acid alpha helix propensity. *Science* **1993**, *260*, 1637–1640. [CrossRef] [PubMed]
66. Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G. iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction. *Front. Immunol.* **2018**, *9*, 1695. [CrossRef]
67. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions. *Front. Immunol.* **2018**, *9*, 1783. [CrossRef] [PubMed]
68. Wang, X.; Yan, R.; Li, J.; Song, J. SOHPRED: A new bioinformatics tool for the characterization and prediction of human S-sulfenylation sites. *Mol. Biosyst.* **2016**, *12*, 2849–2858. [CrossRef]
69. Bhadra, P.; Yan, J.; Li, J.; Fong, S.; Siu, S.W.I. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.* **2018**, *8*, 1697. [CrossRef]
70. Zhang, P.; Tao, L.; Zeng, X.; Qin, C.; Chen, S.Y.; Zhu, F.; Yang, S.Y.; Li, Z.R.; Chen, W.P.; Chen, Y.Z. PROFEAT Update: A Protein Features Web Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks. *J. Mol. Biol.* **2017**, *429*, 416–425. [CrossRef] [PubMed]
71. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest. *Front. Pharmacol.* **2018**, *9*, 276. [CrossRef] [PubMed]
72. Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *J. Proteome Res.* **2018**, *17*, 2715–2726. [CrossRef]
73. Dao, F.Y.; Lv, H.; Wang, F.; Feng, C.Q.; Ding, H.; Chen, W.; Lin, H. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* **2018**. [CrossRef]
74. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
75. Li, Y.; Niu, M.; Zou, Q. ELM-MHC: An Improved MHC Identification Method with Extreme Learning Machine Algorithm. *J. Proteome Res.* **2019**, *18*, 1392–1401. [CrossRef]
76. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
77. Cao, R.; Bhattacharya, D.; Hou, J.; Cheng, J. DeepQA: Improving the estimation of single protein model quality with deep belief networks. *BMC Bioinform.* **2016**, *17*, 495. [CrossRef] [PubMed]
78. Cao, R.; Freitas, C.; Chan, L.; Sun, M.; Jiang, H.; Chen, Z. ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* **2017**, *22*, 1732. [CrossRef]
79. Wei, L.; Su, R.; Wang, B.; Li, X.; Zou, Q.; Gao, X. Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. *Neurocomputing* **2018**, *324*, 3–9. [CrossRef]
80. Malik, A.; Ahmad, S. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. *BMC Struct. Biol.* **2007**, *7*, 1. [CrossRef]
81. Malik, A.; Firoz, A.; Jha, V.; Ahmad, S. PROCARB: A Database of Known and Modelled Carbohydrate-Binding Protein Structures with Sequence-Based Prediction Tools. *Adv. Bioinform.* **2010**, *436036*. [CrossRef]
82. Su, R.; Hu, J.; Zou, Q.; Manavalan, B.; Wei, L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief Bioinform.* **2019**. [CrossRef] [PubMed]