Bioinformatics, 35(16), 2019, 2757–2765
doi: 10.1093/bioinformatics/bty1047
Advance Access Publication Date: 24 December 2018
Original Paper

Sequence analysis

mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation

Balachandran Manavalan[1,†], Shaherin Basith[1,†], Tae Hwan Shin[1,2], Leyi Wei [3,*] and Gwang Lee[1,2,*]

[1]Department of Physiology, Ajou University School of Medicine, Suwon, Republic of Korea, [2]Institute of Molecular Science and Technology, Ajou University, Suwon, Republic of Korea and [3]School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on November 2, 2018; revised on December 5, 2018; editorial decision on December 14, 2018; accepted on December 20, 2018

Abstract

Motivation: Cardiovascular disease is the primary cause of death globally accounting for approximately 17.7 million deaths per year. One of the stakes linked with cardiovascular diseases and other complications is hypertension. Naturally derived bioactive peptides with antihypertensive activities serve as promising alternatives to pharmaceutical drugs. So far, there is no comprehensive analysis, assessment of diverse features and implementation of various machine-learning (ML) algorithms applied for antihypertensive peptide (AHTP) model construction.

Results: In this study, we utilized six different ML algorithms, namely, Adaboost, extremely randomized tree (ERT), gradient boosting (GB), *k*-nearest neighbor, random forest (RF) and support vector machine (SVM) using 51 feature descriptors derived from eight different feature encodings for the prediction of AHTPs. While ERT-based trained models performed consistently better than other algorithms regardless of various feature descriptors, we treated them as baseline predictors, whose predicted probability of AHTPs was further used as input features separately for four different ML-algorithms (ERT, GB, RF and SVM) and developed their corresponding meta-predictors using a two-step feature selection protocol. Subsequently, the integration of four meta-predictors through an ensemble learning approach improved the balanced prediction performance and model robustness on the independent dataset. Upon comparison with existing methods, mAHTPred showed superior performance with an overall improvement of approximately 6–7% in both benchmarking and independent datasets.

Availability and implementation: The user-friendly online prediction tool, mAHTPred is freely accessible at http://thegleelab.org/mAHTPred.

Contact: weileyi@tju.edu.cn or glee@ajou.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Hypertension (HT) known as high blood pressure is the most common global health concern that affects 25% of the population and its occurrence increases with age (Chockalingam *et al.*, 2006). Besides heart-related diseases, HT can also lead to other abnormalities like renal failure, multi-infarct dementia, amputation and HT retinopathies (Varounis *et al.*, 2016). Due to its high prevalence and detrimental effects, it has been necessitated to discover novel drugs and treatments to reduce or eliminate HT-related repercussions. Currently, there are a number of effective drugs available in the market for the treatment of HT, such as alpha- and beta-blockers, angiotensin-converting enzyme (ACE) inhibitors, angiotensin-II receptor blockers, calcium channel blockers, diuretics, peripheral adrenergic inhibitors, renin inhibitors and vasodilators (Puchalska *et al.*, 2015). However, these drugs induce notable side-effects, including cough, dizziness, fatigue, headaches, hyperkalemia, hypotension, impaired taste, increased potassium level, leg edema and skin rashes (Husserl and Messerli, 1981). Therefore, it remains challenging to design and deliver safer drugs for prevention and treatment of HT.

The renin–angiotensin system that plays a pivotal role in regulating arterial pressure is linked with blood pressure control. Renin converts angiotensinogen to the decapeptide angiotensin I, which further undergoes proteolytic cleavage by ACE to biologically active octapeptide, angiotensin II (Dostal and Baker, 1999; Hong *et al.*, 2008). Angiotensin II plays a vital role in vasoconstriction, stimulates aldosterone production and increases sodium and fluid retention. Thus, inhibition of these enzymes aid in reducing blood pressure. Several bioactive peptides that inhibit angiotensin I, ACE and angiotensin II type I receptor in the cardiovascular system help in the prevention and treatment of HT (Hong *et al.*, 2008). These bioactive peptides could be extracted from plant and animal sources, including milk and dairy products, pork, meat, fish, blood, ovalbumin, rice, wheat, potato, cereal, peas, garlic, etc. (Dostal and Baker, 1999). Such identification of bioactive peptides as antihypertensive peptides (AHTPs) led to the implementation of peptides as safe and effective drugs in the treatment of HT (Bhat *et al.*, 2017; Jakala and Vapaatalo, 2010; Majumder and Wu, 2014; Puchalska *et al.*, 2015). Moreover, due to the laborious and time-consuming experimental procedures, it is compelling to develop an effective computational approach for classifying available peptides as AHTP or non-AHTP.

In the recent decade, few computational studies (summarized below) have shown the potency of machine-learning (ML) approaches in AHTP classification. Initially, Wang *et al.* built QSAR models of ACE-inhibitor oligopeptides based on G-scale descriptors using partial least square (PLS) regression method (Wang *et al.*, 2011). The drawback of this method is its applicability for the prediction of inhibitory activity of tiny peptides (i.e. di- and tripeptides) only. In 2015, Kumar *et al.* developed four different model types for predicting AHTPs with varied lengths [i.e. tiny (di- and tripeptides), small (tetra-, penta- and hexapeptides), medium (sizes ranging from 7 to 12) and large peptides (greater than 12 amino acids)] using ML approaches (Kumar *et al.*, 2015b). For tiny peptides, SVM-based regression models were developed using chemical descriptors, and correlations of 0.701 and 0.543 were obtained for di- and tripeptides, respectively. For smaller peptides, SVM-based classification models were built and accuracies of 76.67, 72.04 and 77.39% were attained for tetrapeptides, pentapeptides and hexapeptides, respectively. Similarly, in the case of medium and large peptides, SVM-based classification models were developed using amino acid compositions, and maximum accuracies of 82.61 and 84.21% were obtained. Also, a web-based platform, AHTpin, a web-based platform, was established for predicting, designing and screening of AHTPs. Recently, another paper on AHTPs prediction using ML approaches was published (Win *et al.*, 2018), where the authors developed classification models based on varied combinations of amino acid, dipeptide and pseudo amino acid composition descriptors using random forest (RF) approach and this method showed marginal improvement over AHTpin. Moreover, the feature importance analysis highlighted the significance of Proline and non-polar amino acids at the carboxyl terminal and the importance of short peptides for robust activity as well. Additionally, an online web server, PAAP, was developed for the proposed model.

Although the above-mentioned methods produced encouraging results and stimulated research on AHTP prediction, there are certain drawbacks associated with these approaches which are as follows: (i) only limited features have been utilized by the state-of-the-art methods, further emphasizing that other potential features yet remain to be defined; (ii) exploration of several ML algorithms on the same benchmarking dataset is necessitated and preference of an appropriate algorithm for a specific problem (AHTP prediction) rather than selecting ML algorithm randomly or choice of interest (employed in the existing methods); and (iii) embodiment of redundant features in model development decreases the performance. Thus, to eliminate redundant features and subsequently enhance the prediction performance, feature selection is usually required. However, the above-mentioned methods failed to adopt these strategies. Therefore, novel and competent computational approaches are necessitated to address the mentioned limitations to provoke more accurate models for efficient AHTPs prediction.

Here, we developed mAHTPred, a new meta-predictor for the identification of AHTPs. Firstly, we applied a feature representation learning scheme to extract informative features (51 features based on probabilistic information) from diverse sequence-based descriptors, including amino acid composition (AAC), amino acid index (AAI), binary profile features (BPF), composition-transition-distribution (CTD), dipeptide composition (DPC), other features (OF), overlapping property features (OVP) and twenty-one-bit features (TOB). Secondly, we inputted 51 features separately into four different ML algorithms [extremely randomized tree (ERT), gradient boosting (GB), RF and SVM] and developed their corresponding optimal meta-predictor using a two-step feature selection protocol. Finally, we integrated these four ML-based meta-predictors into an ensemble model for the final prediction. Comparative results with the existing methods on benchmark and independent datasets showed that mAHTPred improvement is significant. To the best of our knowledge, our study is the first meta-based approach in the prediction of AHTPs. Henceforth, we highly anticipate that our work will instigate the development of novel computational approaches and also will facilitate experimentalists in the discovery of novel AHTPs.

2 Materials and methods

The mAHTPred methodology (Fig. 1) consists of five major steps: (i) construction of benchmarking and independent datasets; (ii) feature extraction that covers several aspects of sequence information; (iii) feature representation learning scheme; (iv) construction of a meta-predictor using two-step feature selection strategy [i.e. feature ranking and sequential forward search (SFS)]; and (v) construction of the final model for the classification. Each of these major steps has been detailed in the following sections.

2.1 Construction of benchmarking and independent datasets

To develop a prediction model, we considered the same non-redundant dataset as originally proposed in (Kumar *et al.*, 2015b) that consisted of peptides ranging from dipeptides to larger (>13 amino acid) peptides. In this study, we excluded peptides whose lengths were < 5 amino acid residues due to the difficulty in generating informative features for the shorter sequences. The remaining peptides were considered as the benchmarking dataset. Our final balanced benchmarking dataset constitutes an equal number of AHTPs (913) and non-AHTPs (913). In case of AHTPs, all the sequences are experimentally validated ones derived from the publicly available databases AHTPDB (Kumar *et al.*, 2015a) and BIOPEP (Iwaniak *et al.*, 2016; Minkiewicz *et al.*, 2008). Due to the lack of experimentally validated non-AHTPs, random peptides generated from Swiss-Prot proteins were considered as negative ones. This approach of considering random sequences as negative dataset has been routinely used in peptide-based prediction methods due to the chances of finding random sequences as positive ones are very minimal (Agrawal *et al.*, 2018; Chen *et al.*, 2016; Manavalan *et al.*, 2017, 2018d; Sharma *et al.*, 2013; Usmani *et al.*, 2018a; Wei *et al.*, 2018c).

To evaluate the performance of our method with the existing tools, we constructed a non-redundant independent dataset. Firstly, we extracted experimentally validated AHTPs by manual curation from various literatures (Win *et al.*, 2018; Yi *et al.*, 2018) and databases, including, AHTPDB (Kumar *et al.*, 2015a) and BIOPEP (Iwaniak *et al.*, 2016; Minkiewicz *et al.*, 2008). Furthermore, random peptides generated from the Swiss-Prot were considered as negative samples. Here, the random peptides similar to AHTPs were removed and considered the remaining ones as non-AHTPs. Subsequently, we applied CD-HIT to remove the sequences which shares a sequence identity of >90% in the independent dataset against the sequences in the benchmarking dataset. Finally, we obtained 386 AHTPs and 386 non-AHTPs.

2.2 Feature representation

Depiction of a peptide sequence (P) is as follows:

$$P = R_1R_2R_3 \ldots . R_n \tag{1}$$

where R_1, R_2 and R_3 respectively denote the 1st, 2nd and 3rd residues, respectively, in a peptide P and so on. n denotes the length of the peptide sequence. Each residue (R_i) in a peptide belong to the standard amino acid. To construct an ML model, peptides with diverse-length were formulated as fixed-length feature vectors. Since the feature extraction influences the performance of the prediction model, we exploited various compositions, hybrid features and profiles that includes several facets of sequence information as detailed below:

Amino acid composition (AAC)

AAC is the percentage of standard amino acids with a fixed length of 20 features (Liu, 2017). Formulation of AAC is as follows:

$$\mathrm{AAC(P)} = (V_1,\ V_2,\ V_3,\ \ldots\ldots, V_{20}) \tag{2}$$

where $V_i = \frac{R_i}{n}(i = 1,\ 2,\ 3, \ldots,\ 20)$ is the percentage of the composition with amino acid type i, R_i is the quantity of type i observed in the protein.

Dipeptide composition (DPC)

DPC is the rate of dipeptides normalized by all possible dipeptide combinations with a fixed length of 400 features (Agrawal *et al.*, 2018; Dhanda *et al.*, 2017; Kumar *et al.*, 2018). Formulation of DPC is as follows:

$$\mathrm{DPC(P)} = (V_1,\ V_2, V_3,\ \ldots\ldots, V_{400}) \tag{3}$$

where $V_i = \frac{R_i}{n}(i = 1,\ 2,\ 3, \ldots,\ 400)$ is the percentage of the composition with dipeptide type i and R_i is the quantity of type i appearing in the protein.

Composition-transition-distribution (CTD)

CTD is employed to delineate the global composition of amino acid property (Dubchak *et al.*, 1995) for a given protein or peptide sequence. Standard amino acids are divided into three different clusters, such as hydrophobic, neutral and polar. Composition (C) computes the percentage composition values of the above three different clusters from a given peptide sequence. Transition (T) computes percentage frequency of a specific property of an amino acid progressed by another property. Distribution (D) constitutes five values for each of the three groups and determines the percentage of a target sequence length within which 25, 50, 75 and 100% of the amino acids of a specific property are situated. A more detailed explanation for this calculation can be found in these studies (Li *et al.*, 2006; Zhang *et al.*, 2017). CTD engenders twenty-one features for each physicochemical property. Furthermore, seven different physicochemical properties (charge, hydrophobicity, normalized van der Waals volume, polarity, polarizability, secondary structure and solvent accessibility) yield a sum of 147 features.

Amino acid index (AAI)

The AAIndex database contains various biochemical and physicochemical properties of amino acids (Kawashima *et al.*, 2007). Recently, Saha *et al.* (2012) identified eight high quality AAIs by clustering 566 AAIs present in the AAIndex database, whose accession codes in the AAIndex database are BIOV880101, BLAM930101, CEDJ970104, LIFS790101, MAXF760101, NA KH920108, TSAJ990101 and MIYS990104 (Liu *et al.*, 2015). These high-quality indices are encoded as 160(=20 × 8)-dimensional vectors from the target peptide sequence. However, the average of these eight high-quality AAIs for each amino acid (a 20-dimensional vector) was used as an input feature to minimize the computational time.

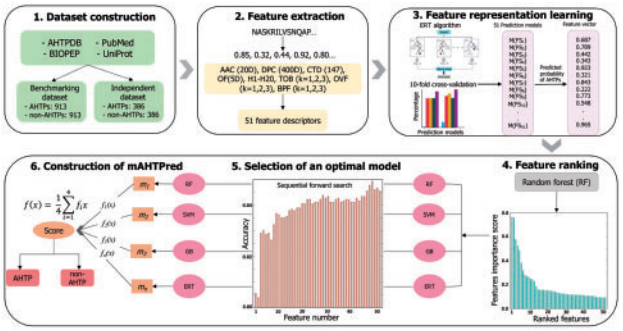

Fig. 1. Overview of the proposed methodology for predicting AHTPs that involved the following steps: (i) construction of benchmarking and independent dataset; (ii) extraction of 8 different feature encodings that characterize those peptides in different ways and generation of 51 feature descriptors; (iii) generation of 51-dimensional feature vector using feature representation learning scheme; (iv) ranking the 51-dimensional feature vector using RF algorithm; (v) generation of the optimal meta-predictor model using sequential forward search; (vi) construction of the final prediction model by integrating four meta-predictors that separates the input into putative AHTPs and non-AHTPs

Other features (OF)

In addition to the above composition, other features are: (i) absolute charge per residue ($|(R+K-D-E)/n-0.03|$); (ii) aliphatic index (i.e. [A + 2.9V + 3.9I + 3.9L]/*n*); (iii) a fraction of turn-forming residues (i.e. [N + G+P + S]/*n*); (iv) molecular weight; and (v) sequence length.

Hybrid features

Generally, hybrid features tend to perform better than individual composition because it contains multiple information from the sequence (Manavalan *et al.*, 2018a, c, d). Hence, we generated hybrid features using a linear combination of five compositions (AAC, DPC, CTD, AAI and OF). Supplementary Table S1 shows the 20 hybrid features employed in this study with various possible combinations covering different perspectives of sequential information.

Binary profile (BPF)

The binary encoding of amino acids converts each amino acid into a 20-dimensional vector. Every amino acid type of 20 different standard amino acids is deciphered with the following feature vector 0/1 (Nagpal *et al.*, 2017; Usmani *et al.*, 2018a, b; Vens *et al.*, 2011). For example, the first amino acid type A is deciphered as b(A) = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), the second amino acid type C is deciphered as b(C) = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) and so on. Consequently, for a given peptide sequence *P*, its N- or C-terminus with a length of *k* amino acids was deciphered as follows:

$$BPF(k) = [b(p_1),\ b(p_2), \ldots, b(p_k)] \tag{4}$$

BPF(*k*) dimension is $20 \times k$, where *k* values are assigned as 3, 4 and 5 at N- and C-terminus, which resulted as follows: BPFN3, BPFN4, BPFN5, BPFC3, BPFC4 and BPFC5. In addition to this, we also generated BPFN3-BPFC3 (BPFNC3), BPFN4-BPFC4 (BPFNC4) and BPFN5-BPFC5 (BPFNC5).

Overlapping property (OVP) features

Based on the physicochemical properties, the standard amino acids are divided in to ten groups (Supplementary Table S2). Due to the possibility of two or more physicochemical properties overlap by a specific amino acid type, there is a chance of overlapping a different group (Govindan and Nair, 2011). To show the relationship of varied properties, we computed a 10-bit vector comprised of 0/1 to depict each amino acid of a given peptide. If a residue of the peptide belongs to each property, the parameter will be set to 1, else 0 (Wei *et al.*, 2018c). From Supplementary Table S2, the amino acid type A is deciphered as b(A) = (0, 0, 0, 0, 1, 0, 1, 0, 1, 0), b(C) = (0, 0, 0, 1, 1, 0, 1, 0, 1, 0) and so on. Consequently, for a given peptide sequence *P*, its N- or C-terminus with length of *k* amino acids was deciphered as follows:

$$OVP(k) = [b(p_1),\ b(p_2), \ldots, b(p_k)] \tag{5}$$

OVP(*k*) dimension is $10 \times k$, where *k* values are assigned as 3, 4 and 5 at N- and C -terminus, which resulted as follows: OVPN3, OVPN4, OVPN5, OVPC3, OVPC4 and OVPC5. In addition to this, we also generated OVPN3-OVPC3 (OVPNC3), OVPN4-OVPC4 (OVPNC4) and OVPN5-OVPC5 (OVPNC5).

Twenty-one-bit (TOB) features

TOB features considers seven physicochemical properties, including charge, hydrophobicity, normalized Van der Waals volume, polarizability, polarity, secondary structure and solvent accessibility (Dou *et al.*, 2014). Supplementary Table S3 shows the classification of amino acid residue into seven physicochemical properties, where any two groups are not overlapped. Similar to OVP deciphering, each residue of the peptide *P* is deciphered as a 21-bit vector composed of 0/1, where the position of each bit position is set to 1 if the amino acid fits in the corresponding group, else 0 (Wei *et al.*, 2018b, c). TOB dimensionality is $21 \times k$, where *k* values are assigned as 3, 4 and 5 at N- and C-terminus, which resulted as follows: TOBN3, TOBN4, TOBN5, TOBC3, TOBC4 and TOBC5. In addition to this, we also generated TOBN3-TOBC3 (TOBNC3), TOBN4-TOBC4 (TOBNC4) and TOBN5-TOBC5 (TOBNC5).

2.3 Feature representation learning scheme

Recently, Wei *et al.* reported a novel feature learning scheme that was successfully applied in various prediction problems, including anticancer peptide (Wei *et al.*, 2018c), cell penetrating peptide (Qiang *et al.*, 2018b) and Quorum sensing peptide predictions (Wei *et al.*, 2018b). In this study, we followed a similar protocol and the steps are as follows:

Step 1. Construction of an initial feature Pool

As mentioned in the previous sections, we extracted eight feature encoding schemes were obtained based on composition, profiles and physicochemical properties, including AAC, AAI, BPF, CTD, DPC, OF, OVP and TOB. Hybrid features contain 20 different feature set (Supplementary Table S1), based on a different combination of five feature encodings, including AAC, DPC, CTD, AAI and OF. In case of BPF, OVP and TOB, the value of *k* was set in the range of 3–5. Since the minimal sequence length of the peptide in our dataset is five residues, we cannot use the value of $k > 5$. Furthermore, we considered N-terminal residues, C-terminal residues and a combination of N- and C-terminal residues, which led to 27 (=9 feature set × 3 encodings) feature set. In total, we generated 51 feature set (FS) based on the eight feature encodings that are listed in Supplementary Table S4. For clarity, the *j*th feature set is represented as FS_j ($j = 1, 2, 3, \ldots, 51$).

Step 2. Construction of feature learning model

For each FS_j ($j = 1, 2, 3, \ldots, 51$), we developed their corresponding ERT-based prediction model, represented as $M(FS_j)$, using benchmarking dataset and 10-fold cross-validation (CV). Acknowledging that running 10-fold CV with random partitioning of benchmarking dataset might yield biased ML parameters, hence, we re-run 10-fold CV for additional five times and considered median ML parameters as the optimal value. Finally, we obtained 51 prediction models and considered them as the baseline model.

Step 3. Learning a new feature vector to construct a Meta predictor

For a given peptide *P*, we used each baseline model ($M(FS_j)$) to determine its probability value of AHTPs, whose value lies in the range of 0-1. The predicted probability value by each model was subsequently used as a feature. In our experiment, predicted probability value ≥ 0.5 belongs to AHTPs, else non-AHTPs. To this end, the sequence *P* is deciphered with a new feature vector (FV) by joining all features produced by 51 models, which is represented as:

$$FV_{ERT}(P) = \ Y(P,\ M(FS_1)),\ Y(P,\ M(FS_1)), \ldots Y(P,\ M(FS_{51})) \tag{6}$$

where $FV_{ERT}(P)$ is the feature vector for a given peptide sequence *P*. $Y(P,\ M(FS_i))$ is the prediction probability of each model for the sequence *P*.

2.4 Construction of meta-predictor

All the features generated in step 3 of feature representation learning scheme (Eq. 6) were subsequently provided as an input discretely to four different ML algorithms (ERT, GB, RF and SVM) and their

corresponding optimal meta-predictor was established using a two-step feature selection strategy. Detailed description of feature selection strategy is mentioned below:

Feature selection

In general, biological datasets are represented as higher dimensional features, which led to decrease the algorithm speed and poor prediction performance (Liu *et al.*, 2017). However, feature selection procedure plays an important role to overcome the above limitations, which is regarded as a potent step in ML-based model development. To enhance the feature representation capability and determine the subset of optimal features from the original 51 features (Eq. 6) which contribute to the appropriate classification of AHTPs or non-AHTPs, a new two-step feature selection strategy was utilized. Remarkably, the two-step feature selection protocol utilized here is similar to the one employed in our recent research (Manavalan and Lee, 2017, 2018b, c, d). In our previous protocol, features were ranked according to the variable or feature importance scores (FISs) using the RF algorithm in the first step, and feature subsets were selected manually in the second step based on the FISs. It is of noteworthy that the first step is same with our previous protocol. But, a SFS was utilized in the second step to select the optimal feature subset (Basith *et al.*, 2018), rather than employing manual feature subset selection.

A given set of features was provided as an input to the RF algorithm and 10-fold CV was performed. For each round of CV, we built 1000 trees were built utilizing a *mtry* range from 1 to 50. To rank the features, average FISs from all the trees were employed.

$$D = [F1,\ F2,\ F3, \ldots\ldots,\ FN]^T \quad (7)$$

where F1 is the first feature with the maximum FIS; F2 is the second feature with the second maximum FIS; F3 is the third feature with the third maximum FIS and so on; N and T are respectively total number of features and the transpose operator.

In the next step, SFS was utilized to identify and select the optimal features from a ranked feature set based on the following steps: (i) The first feature subset contained only the first feature in the ranked set D. The second feature subset contained the first and the second feature in D, and so on. Lastly, we obtained N feature subsets; (ii) All the N feature subsets were inputted to four different ML algorithms (ERT, GB, RF and SVM) for the development of their corresponding prediction model using a 10-fold CV test (Supplementary Material). Certainly, the best performance in terms of accuracy produced by the feature subset was regarded as the optimal feature set.

2.5 Implemented machine learning algorithms for model development

mAHTPred utilizes four different ML algorithms such as ERT, GB, RF and SVM, which were implemented using the Scikit-Learn package (v0.18) (Abraham *et al.*, 2014). Details and utility of these methods in this study along with the evaluation metrics are provided in the Supplementary Material.

3 Results and discussions

3.1 Impact of various classifiers on feature learning models

In this study, we generated 51 feature descriptors as described in Section 2.3 using eight different feature encodings and varied the parameters for only three feature encodings (Supplementary Table S4). Using these feature descriptors, we examined the predictive performance of six commonly-used ML algorithms or classifiers, namely AB, ERT, GB, *k*-NN, RF and SVM, by performing 10-fold CV. In total, we obtained 306 prediction models with six ML algorithms using 51 feature descriptors, whose performances are summarized in Supplementary Tables S5–S10. Apart from ERT and GB, the remaining four ML algorithms achieved their best performances using different feature descriptors, we observed that RF, ERT, GB, AB and *k*-NN achieved their corresponding maximum accuracies of 82.0, 82.3, 80.9, 76.8 and 75.7% using CTD, H11 (a linear combination of AAC, AAI, and DPC), H8 (AAI and DPC), H11, H15 (DPC, CTD, and OF) and H6 (AAC and AAI) feature descriptors, respectively. These results show that the performance of existing feature descriptors is greatly influenced by the utilized classifiers. Surprisingly, the top 10 models (Supplementary Tables S5–S10) for each classifier showed similar performances, and none of the feature descriptors provided a significant performance improvement as we expected. Furthermore, comparisons of the best feature descriptors among six predictors showed that H11 and ERT classifier was somewhat better than RF and GB, and remarkably better than AB and *k*-NN in terms of accuracy (0.3–6.6%) and MCC (0.3–14%). Based on our analysis, we conclude that the predictive model trained with ERT classifier and H11 descriptor has relatively high discriminative power to classify AHTPs from non-AHTPs.

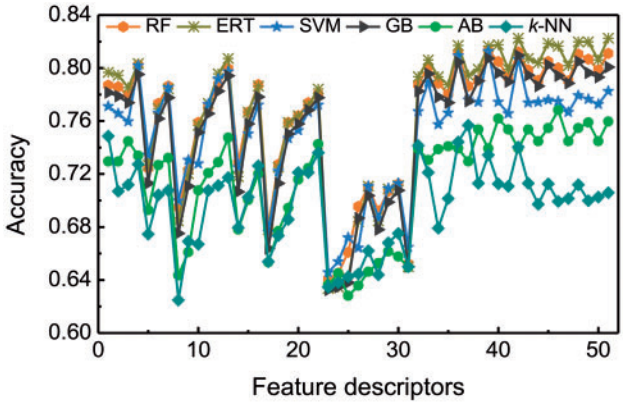

Fig. 2. Performance of various classifiers in distinguishing between AHTPs and non-AHTPs with respect to 51 feature descriptors

Additionally, the effectiveness of the classifiers in predicting AHTPs were explored. Figure 2 shows the performance of six classifiers with respect to 51 feature descriptors. For each classifier, the performances in terms of accuracy was fluctuating and we did not observe any stable performance between feature descriptors. Among various descriptors, TOB feature encoding with varying parameters performed poorer than other descriptors regardless of the ML algorithm, indicating that it has a less discriminative power. Overall, we observed that the accuracy of the feature descriptors using the ERT classifier is generally higher than other classifiers, thus demonstrating its superiority. Therefore, we selected these models for learning feature representation. In case of other classifiers, RF, GB and SVM seem to be competitive with each other, while AB and *k*-NN showed worst performances among the compared classifiers.

3.2 Construction of meta-predictors using two-step feature selection strategy

Since AB and *k*-NN showed worst performances among the six classifiers in AHTPs prediction, we excluded these methods from further analysis and included only the remaining four methods, namely SVM, ERT, RF and GB for the construction of meta-predictor.

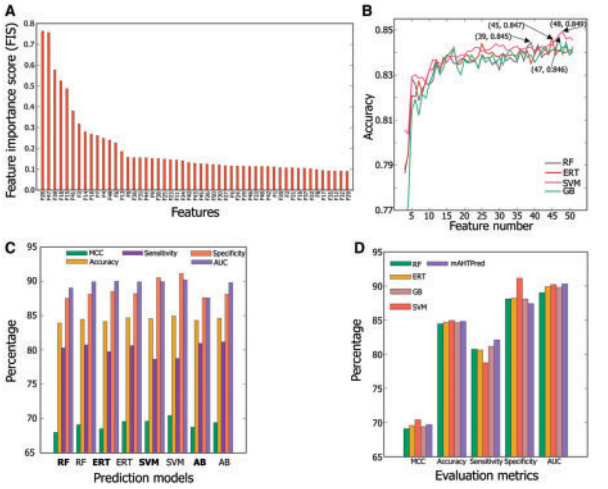

Fig. 3. Feature selection and the final model construction. (**A**) Ranking of 51-dimensional vector according to the feature importance score. (**B**) SFS curve for discriminating AHTPs and non-AHTPs. The maximum accuracy (i.e. SFS peak) obtained in 10-fold CV for the four different methods, namely ERT, GB, RF and SVM is shown in arrow. (**C**) Comparison of the optimal model (normal font) with respect to the control (bold) (i.e. using all the features). (**D**) Comparison of mAHTPred with the individual ML-based prediction model

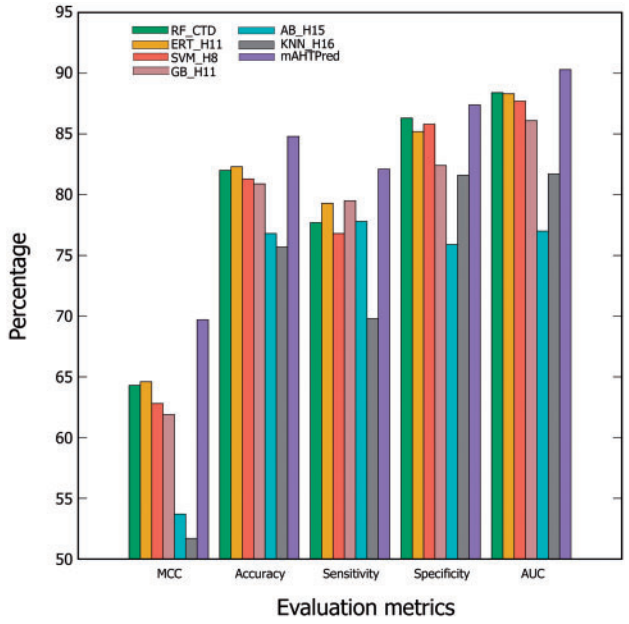

Fig. 4. Performance of mAHTPred and the base-line models. The Performance comparison between mAHTPred and the base-line models in terms of MCC, accuracy, sensitivity, specificity and AUC

In general, meta-predictor consider output from a variety of individual algorithms under the assumption that combined methods provide more accurate predictions than the single method. Here, ERT-based predicted AHTP probability values derived from 51 learnt features were provided as inputs to four ML algorithms and their corresponding optimal meta-predictor models were developed using a two-step feature selection protocol. Primarily, we sorted 51 learnt features and ranked them in accordance with FIS produced by RF algorithm (Fig. 3A and Supplementary Table S11). As shown, it is evident that 28th [H10 (AAI and CTD)] and 47th [H14 (AAI, DPC and CTD)] features which have a FIS of approximately 0.76 appeared to be essential and thus proved most potent for classification. Subsequently, we added the features in succession from the ranked features to the previous ones and their respective prediction models were built. Figure 3B shows the performance in terms of accuracy corresponding to feature number using SFS, where performance of each method steadily increased in parallel with feature number until 15 and it remained stable thereafter. A model with the highest accuracy was selected as the best optimal model, whose corresponding features were regarded as the optimal feature set. Unlike the performances in feature learning, the selected best model of SVM, RF, ERT and GB produced a similar performance of approximately 85.0% with a larger feature number of 48, 39, 47 and 45, respectively. Although we expected a significant improvement from the two-step feature selection strategy as reported in previous studies (An *et al.*, 2016; Dao *et al.*, 2018; Lai *et al.*, 2017; Qiang *et al.*, 2018a, b; Song *et al.*, 2018; Zhang *et al.*, 2018), the improvement of the optimal four models on an average was very marginal (~0.5%) when compared to the control (using all the features) (Fig. 3C). This might be due in part to the optimal feature size (on an average ~45 features), which is almost similar to the control (51 features).

3.3 Construction of mAHTPred

Since the performances of four meta-predictor models are similar, we integrated these models into an ensemble model called mAHTPred, which are as follows:

$$E = RF \forall SVM \forall ERT \forall GB$$

where E and $\forall$ refers to the ensemble model and fusion operator, respectively. Subsequent to fusion, the average probability cut-off values corresponding to the accuracy using grid search to define the class (as AHTPs or non-AHTPs) were optimized. Best performance was observed with a 0.44 cut-off, henceforth we fixed it as an optimal cut-off. Therefore,

$$S = \begin{cases} AHTPs, & if\ E \geq 0.44 \\ non-AHTPs, & else \end{cases}$$

Although, we observed that mAHTPred performance in terms of MCC and accuracy is similar with other methods, but it achieved more balanced prediction results (Fig. 3D). Notably, the four meta-predictor models achieved a very high specificity and comparatively lower sensitivity with an average difference of ~9% (i.e. Specificity-Sensitivity), however, mAHTPred achieved 4% lesser than the above methods.

To show the advantage of meta-predictor, we compared the performance of mAHTPred with the best models from each of the six different ML algorithms obtained from feature learning (Supplementary Tables S5–S10). Figure 4 shows that MCC and accuracy of mAHTPred were respectively 5.0–18.0% and 2.5–9.0% higher than the single models. Furthermore, we compared the AUC between mAHTPred and other methods and computed *P*-value using two-tailed t-test (Hanley and McNeil, 1982). Using a *P*-value threshold of 0.05, mAHTPred significantly outperformed all single models. While this approach has been quite commonly applied for protein structure (Bujnicki *et al.*, 2001) and peptide function predictions (Wei *et al.*, 2018b, c), however, it is of noteworthy that this is the first illustration where meta-predictor method has been employed for AHTPs prediction.

3.4 Feature selection analysis

To understand the effectiveness of our features, we computed T-distributed Stochastic Neighbor Embedding (t-SNE) for positive and

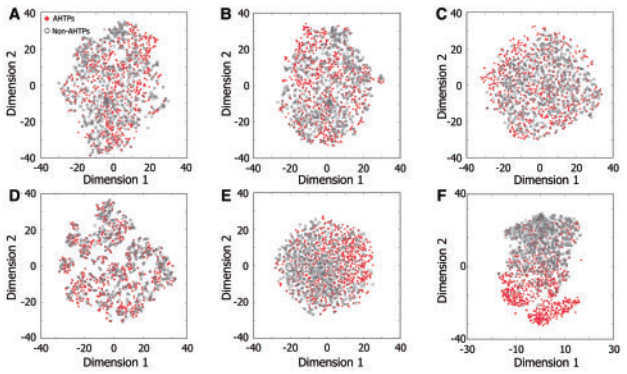

Fig. 5. t-SNE distribution of AHTPs and non-AHTPs using 51-dimensional vector and the top five individual descriptors. (**A–F**) are the distribution of H10, H4, OVPN5, TOBNC4, BPFNC4 and 51-dimensional vectors, respectively

negative samples of the 51-dimensional vector and compared with the top five individual feature descriptors which have been ranked by RF algorithm. Figure 5A–E shows that positive and negative samples of five feature descriptors are distributed differentially in the feature space (A–E). Conversely, we observed a clear distinction between positive and negative samples for 51-dimensional vector, although few samples appear overlaid (Fig. 5F). These results demonstrate that AHTPs and non-AHTPs present in 51-dimensional vector could be easily differentiated when compared to the other feature space, thus enhancing the performance. Our feature selection protocol could be proven effective due to the following reasons: (i) no tuning of the parameters is necessary for datasets as commonly done by most of the existing descriptors; (ii) it can be easily scalable for both peptide and protein feature representations; and (iii) easy transformation from high-dimensional feature space into low-dimensional one is possible, thus leading to the expedition of prediction process and extending its applicability to genome-wide predictions too.

3.5 Performance evaluation of mAHTPred with other predictors on benchmarking dataset

To evaluate the performance of mAHTPred, our computational protocol was compared with two exiting methods available in the literature, namely AHTpin and PAAP. It is to be noted that AHTpin has two prediction models (the first one is based on AAC and the second one is based on atomic composition) and we used both in our analysis. The rationale for considering these two methods in our analysis are as follows: (i) the authors trained and validated their prediction models using the same benchmarking dataset as presented in this study and (ii) these methods have been reported to demonstrate excellent performance in AHTPs identification. As shown in Table 1, we observed that among the compared predictors, mAHTPred demonstrated the best performance in terms of MCC, accuracy, sensitivity and specificity of 0.697, 84.8, 82.1 and 87.4%, respectively. Indeed, MCC and accuracy of mAHTPred were respectively 11.2–13 and 5.7–6.3% higher than existing methods, thus demonstrating the superiority of our proposed protocol. Furthermore, to evaluate the generalization, robustness and practical applicability of our method, we evaluated the performances of all these methods on independent dataset.

Table 1. Performance comparison of between our proposed method and the state-of-the-art methods for predicting AHTPs based on the benchmarking dataset

Methods	MCC	Acc	Sn	Sp	AUC
mAHTPred	0.697	0.848	0.821	0.874	0.903
PAAP	0.585	0.791	0.865	0.780	NA
AHTpin_AAC	0.567	0.785	0.777	0.793	NA
AHTpin_ATC	0.573	0.785	0.783	0.787	NA

Note: First column represents the method name employed in this study. The second, third, fourth, fifth and the sixth columns, respectively, represent the MCC, Acc: accuracy, Sn: sensitivity, Sp: specificity and AUC. NA: not available.

Table 2. Performance comparison of between our proposed method and the state-of-the-art methods for predicting AHTPs based on the independent dataset

Methods	MCC	ACC	Sn	Sp	AUC	*P*-value (AUC)
mAHTPred	0.767	0.883	0.894	0.873	0.951	—
AHTpin_ATC	0.641	0.820	0.798	0.842	0.888	**0.000015**
AHTpin_AAC	0.601	0.800	0.821	0.780	0.852	**<0.000001**

Note: First column represents the method name employed in this study. The second, third, fourth, fifth and the sixth columns, respectively, represent the MCC, ACC: accuracy, Sn: sensitivity, SP: specificity and AUC. The last column represents the pairwise comparison of AUCs between mAHTPred and the other methods using a two-tailed t-test. $P < 0.01$ indicates a statistically meaning full difference between the mAHTPred and the existing method (shown in bold).

3.6 Performance evaluation of mAHTPred with other predictors on the independent dataset

To assess the robustness of mAHTPred, its performance using independent dataset was compared with AHTpin only because the other reported method's PAAP webserver was not functional during our manuscript preparation. As shown in Table 2, mAHTPred showed the best performance in terms of MCC, accuracy, sensitivity, specificity and AUC of 0.767, 88.3, 89.4, 87.3 and 95.1%, respectively. Explicitly, MCC and accuracy of mAHTPred were approximately 6.3–8.3% and 12.6–16.6% higher than the existing method, thus demonstrating the superiority of our proposed predictor. Furthermore, we plotted ROC curve (Supplementary Fig. S1) which provide a comprehensive performance comparison between mAHTPred and other method. mAHTPred significantly outperformed the existing method using a *P*-value threshold of 0.01, thus demonstrating that our model was indeed a robust one in the accurate prediction of AHTPs.

The consistent performance of mAHTPred on both benchmarking and independent dataset suggest that our method could accurately identify AHTPs from unknown peptides. The rationale for an improved performance of mAHTPred over the existing method are as follows: (i) our feature learning model uses an enlarged set of informative sequence-based features, including residue composition, sequence local-order information, physicochemical properties and residue position specific information. (ii) Our model uses the probability of predicted AHTPs from the original feature descriptors, which significantly reduced the high-dimensional complex feature space into a low-dimensional and more informative one; and (iii) our final ensemble integrates four meta-predictors, which further led to a more stable performance.

4 Conclusion

Hypertension is linked to several diseases including cancer, cardiovascular diseases, renal diseases and other complications. Naturally derived bioactive peptides with antihypertensive activities serve as promising alternatives to pharmaceutical drugs. Therefore, accurate identification of AHTPs from provided sequence information seems as one of the challenging tasks in bioinformatics. Although there have been few computational methods to predict AHTPs, a systematic comprehensive assessment of informative features, the effectiveness of ML algorithms, and their potential integration have been lacking. In this study, we conducted a comprehensive analysis of 51 feature descriptors using six different ML algorithms for the computational identification of AHTPs. In order to develop a high efficiency predictor, we implemented the following protocol: (i) we applied a feature representation learning and extracted more informative features using ERT algorithm; (ii) the predicted probability of AHTPs were utilized as an input to SVM, RF, ERT and GB classifiers, and their related optimal meta-predictors were built using a two-step feature selection strategy; and (iii) a combination of these four meta-predictors into ensemble strategies produced a more stable performance. Our analysis highlighted that mAHTPred showed consistently better performance on both benchmarking and independent datasets, indicating that the proposed method is more pragmatic and idealistic for the prediction of AHTPs. Additionally, we made our method available in the form of a free web server for the easy accessibility and utility to a wider research community. We expect that mAHTPred will be a powerful bioinformatics tool for identifying new potential AHTPs in an effective and economical manner. Moreover, our proposed computational framework will not only be applicable to AHTPs but could be further extended to other peptide sequence-based predictors (e.g. cell-penetrating peptides, antimicrobial peptides and antibacterial peptides), as well as other bioinformatics fields (Cui *et al.*, 2018; McDermaid *et al.*, 2018). Additionally, it can be expected that integrating other informative features, such as conserved motif features, might further improve the performance of sequence-based predictors (Ma *et al.*, 2013; Yang *et al.*, 2017).

Acknowledgements

The authors would like to thank Ms. Dae Yeon Lee and Ms. Saraswathi Nithyanantham for their support in dataset preparation.

Funding

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science, and Technology [2018R1D1A1B07049572 and 2018R1D1A1B07049494], ICT & Future Planning [2016M3C7A1904392], the National Natural Science Foundation of China (No. 61701340) and the Natural Science Foundation of Tianjin city (No. 18JCQNJC00500).

Conflict of Interest: none declared.

References

Abraham,A. *et al.* (2014) Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.*, **8**, 14.

Agrawal,P. *et al.* (2018) In silico approach for prediction of antifungal peptides. *Front. Microbiol.*, **9**, 323.

An,Y. *et al.* (2016) Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. *Brief. Bioinform.*, **19**, 148–161.

Basith,S. *et al.* (2018) iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.*, **16**, 412–420.

Bhat,Z.F. *et al.* (2017) Antihypertensive peptides of animal origin: a review. *Crit. Rev. Food Sci. Nutr.*, **57**, 566–578.

Bujnicki,J.M. *et al.* (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.

Chen,W. *et al.* (2016) iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget*, **7**, 16895–16909.

Chockalingam,A. *et al.* (2006) Worldwide epidemic of hypertension. *Can. J. Cardiol.*, **22**, 553–555.

Cui,X. *et al.* (2018) UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. *Chemometr. Intell. Lab. Syst.*, **184**, 28–43.

Dao,F.Y. *et al.* (2018) Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics*. in press.

Dhanda,S.K. *et al.* (2017) Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. *Brief. Bioinform.*, **18**, 467–478.

Dostal,D.E. and Baker,K.M. (1999) The cardiac renin–angiotensin system: conceptual, or a regulator of cardiac function? *Circ. Res.*, **85**, 643–650.

Dou,Y. *et al.* (2014) PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids*, **46**, 1459–1469.

Dubchak,I. *et al.* (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA*, **92**, 8700–8704.

Govindan,G. and Nair,A.S. (2011) Composition, Transition and Distribution (CTD)—a dynamic feature for predictions based on hierarchical structure of cellular sorting. In *India Conference (INDICON), 2011 Annual IEEE*. IEEE. pp. 1–6.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hong,F. *et al.* (2008) The antihypertensive effect of peptides: a novel alternative to drugs? *Peptides*, **29**, 1062–1071.

Husserl,F.E. and Messerli,F.H. (1981) Adverse effects of antihypertensive drugs. *Drugs*, **22**, 188–210.

Iwaniak,A. *et al.* (2016) BIOPEP database of sensory peptides and amino acids. *Food Res. Int.*, **85**, 155–161.

Jakala,P. and Vapaatalo,H. (2010) Antihypertensive peptides from milk proteins. *Pharmaceuticals*, **3**, 251–272.

Kawashima,S. *et al.* (2007) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.

Kumar,R. *et al.* (2015a) AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic Acids Res.*, **43**, D956–D962. (Database issue)

Kumar,R. *et al.* (2015b) An in silico platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep.*, **5**, 12512.

Kumar,V. *et al.* (2018) Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. *Front. Microbiol.*, **9**, 725.

Lai,H.-Y. *et al.* (2017) Sequence-based predictive modeling to identify cancerlectins. *Oncotarget*, **8**, 28169.

Li,Z.R. *et al.* (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **34**, W32–W37.

Liu,B. (2017) BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.*, in press.

Liu,B. *et al.* (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–W71.

Liu,B. *et al.* (2017) Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget*, **8**, 13338–13343.

Ma,Q. *et al.* (2013) An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. *Bioinformatics*, **29**, 2261–2268.

Majumder,K. and Wu,J. (2014) Molecular targets of antihypertensive peptides: understanding the mechanisms of action based on the pathophysiology of hypertension. *Int. J. Mol. Sci.*, **16**, 256–283.

Manavalan,B. *et al.* (2017) MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget*, **8**, 77121–77136.

Manavalan,B. and Lee,J. (2017) SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics*, **33**, 2496–2503.

Manavalan,B. *et al.* (2018a) AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.*, **9**, 276.

Manavalan,B. *et al.* (2018b) DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget*, **9**, 1944–1956.

Manavalan,B. *et al.* (2018c) PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.*, **9**, 476.

Manavalan,B. *et al.* (2018d) Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.*, **17**, 2715–2726.

McDermaid,A. *et al.* (2018) A new machine learning-based framework for mapping uncertainty analysis in RNA-Seq read alignment and gene expression estimation. *Front. Genet.*, **9**, 313.

Minkiewicz,P. *et al.* (2008) BIOPEP database and other programs for processing bioactive peptide sequences. *J. AOAC Int.*, **91**, 965–980.

Nagpal,G. *et al.* (2017) Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci. Rep.*, **7**, 42851.

Puchalska,P. *et al.* (2015) Isolation and characterization of peptides with antihypertensive activity in foodstuffs. *Crit. Rev. Food Sci. Nutr.*, **55**, 521–551.

Qiang,X. *et al.* (2018a) M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. *Front. Genet.*, **9**, 495.

Qiang,X. *et al.* (2018b) CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief. Bioinform.*, in press.

Saha,I. *et al.* (2012) Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids*, **43**, 583–594.

Sharma,A. *et al.* (2013) Computational approach for designing tumor homing peptides. *Sci. Rep.*, **3**, 1607.

Song,J. *et al.* (2018) iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.*, in press.

Usmani,S.S. *et al.* (2018a) Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features. *Front. Pharmacol.*, **9**, 954.

Usmani,S.S. *et al.* (2018b) In silico tools and databases for designing peptide-based vaccine and drugs. *Adv. Protein. Chem. Struct. Biol.*, **112**, 221–263.

Varounis,C. *et al.* (2016) Cardiovascular hypertensive crisis: recent evidence and review of the literature. *Front. Cardiovasc. Med.*, **3**, 51.

Vens,C. *et al.* (2011) Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, **27**, 1231–1238.

Wang,X. *et al.* (2011) QSAR study on angiotensin-converting enzyme inhibitor oligopeptides based on a novel set of sequence information descriptors. *J. Mol. Model.*, **17**, 1599–1606.

Wei,L. *et al.* (2018a) M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids*, **12**, 635–644.

Wei,L. *et al.* (2018b) Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Brief. Bioinform.*, in press.

Wei,L. *et al.* (2018c) ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics.*, **34**, 4007–4016.

Win,T.S. *et al.* (2018) PAAP: a web server for predicting antihypertensive activity of peptides. *Future Med. Chem.*, **10**, 1749–1767.

Yang,J. *et al.* (2017) DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics*, **33**, 2586–2588.

Yi,Y. *et al.* (2018) High throughput identification of antihypertensive peptides from fish proteome datasets. *Mar Drugs*, **16**, 365.

Zhang,P. *et al.* (2017) PROFEAT update: a protein features web server with added facility to compute network descriptors for studying omics-derived networks. *J. Mol. Biol.*, **429**, 416–425.

Zhang,Y. *et al.* (2018) Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.*, in press.