Computational Biology and Chemistry 80 (2019) 441–451

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

THPep: A machine learning-based approach for predicting tumor homing peptides

Watshara Shoombuatong[a,*], Nalini Schaduangrat[a], Reny Pratiwi[a,b], Chanin Nantasenamat[a,*]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Department of Medical Laboratory Technology, Faculty of Health Science, Setia Budi University, Surakarta 57127, Indonesia

ARTICLE INFO

Keywords:
Tumor homing peptide
Therapeutic peptide
Classification
Machine learning
Random forest

ABSTRACT

In the present era, a major drawback of current anti-cancer drugs is the lack of satisfactory specificity towards tumor cells. Despite the presence of several therapies against cancer, tumor homing peptides are gaining importance as therapeutic agents. In this regard, the huge number of therapeutic peptides generated in recent years, demands the need to develop an effective and interpretable computational model for rapidly, effectively and automatically predicting tumor homing peptides. Therefore, a sequence-based approach referred herein as THPep has been developed to predict and analyze tumor homing peptides by using an interpretable random forest classifier in concomitant with amino acid composition, dipeptide composition and pseudo amino acid composition. An overall accuracy and Matthews correlation coefficient of 90.13% and 0.76, respectively, were achieved from the independent test set on an objective benchmark dataset. Upon comparison, it was found that THPep was superior to the existing method and holds high potential as a useful tool for predicting tumor homing peptides. For the convenience of experimental scientists, a web server for this proposed method is provided publicly at http://codes.bio/thpep/.

1. Introduction

The burden of cancer has become a major cause of concern worldwide owing to its associated high morbidity and mortality. According to the World Health Organization (WHO) and the National Cancer Institute (NCI), more than 10 million new cases are reported each year and 1 in 6 of the global deaths reported is due to cancer (N. I. o. Health, 2017; W. H. Organization, 2017). Despite the progression in cancer drug development, the lack of specificity of chemotherapeutic drugs towards tumor cells and its toxicity to normal cells still represent key challenges in this field (Mäe et al., 2009; Myrberg et al., 2008). In an attempt to discover a more potent and specific therapy for tumor, the research community has focused their efforts in exploiting peptides as a tool to improve the targeted-drug delivery system. Over the last decade, three distinct classes of peptide-mediated drug delivery systems have been developed including, homing peptides (HPs), peptide linked to cell-penetrating peptides (CPPs), and cell-penetrating homing peptides (CPHPs). HPs functions by delivering cargo (i.e. drugs) on to the cell surface, while CPPs facilitates the cargo internalization, and CPHPs enables cargo internalization without support form an external agent (Svensen et al., 2012). Herein, we have focused on a tumor homing peptide (THP), which is a short sequence of peptides with the specific ability to recognize and home to (i.e. reach out) to tumor cells or tumor vasculature (Laakkonen and Vuorinen, 2010).

Since the discovery of in vivo phage display technology and in vitro screening of synthetic peptide libraries, a large number of THPs have been identified (Svensen et al., 2012). The first generation of THPs contain common motifs like RGD (Arg-Gly-Asp) and NGR (Asn-Gly-Arg) (Laakkonen and Vuorinen, 2010). The RGD peptide binds specifically to α-integrin receptors, while the NGR-containing peptides bind to aminopeptidase N receptors, which are expressed in tumor vasculature (Pasqualini et al., 1997, 2000). Because of their short length, usually between 3–15 amino acids, these peptides are more favorable to travel across the body than other anti-tumor drugs as show in Fig. 1. In addition, there is a vast literature stating that THPs exhibited well-defined specificity towards receptors on target organs which may be part of the tumor vasculature or lymphatic system. Furthermore, these peptides are relatively non-immunogenic and generate low production costs as compared to other therapeutic vehicles (Svensen et al., 2012; Laakkonen and Vuorinen, 2010; Gautam et al., 2014).

Although experimental approach is an objective method to identify THPs, it is time-consuming. Particularly, with a huge number of therapeutic peptides generated in recent years, it is highly desirable to develop a computational model to discriminate THPs from non-THPs.

* Corresponding authors.
E-mail addresses: watshara.sho@mahidol.ac.th (W. Shoombuatong), chanin.nan@mahidol.edu (C. Nantasenamat).

https://doi.org/10.1016/j.compbiolchem.2019.05.008
Received 6 December 2018; Received in revised form 18 April 2019; Accepted 17 May 2019
Available online 24 May 2019

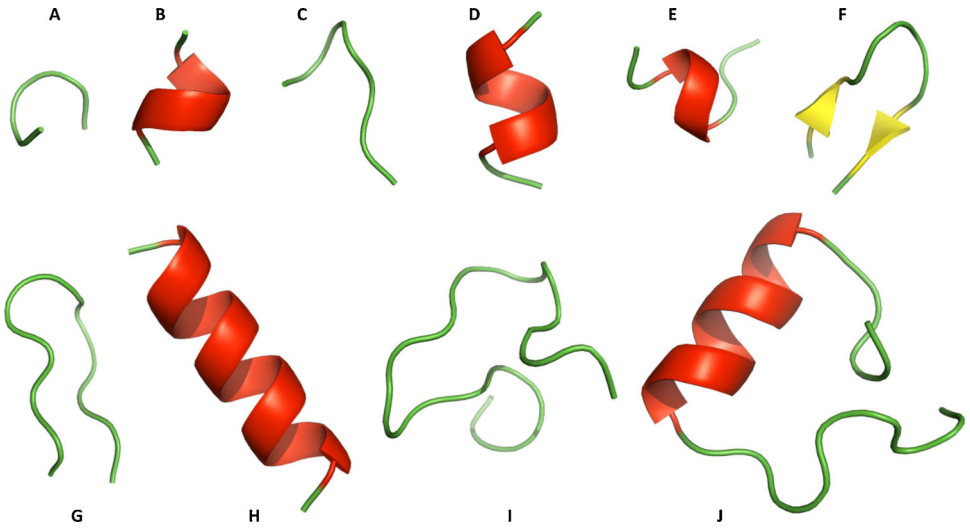

Fig. 1. Three-dimensional structure of THPs. Panels A–F are THPs with 5–10 amino acid residues while panels G–J are THPs larger than 10 amino acid residues. Peptide structures were constructed by PEP-FOLD server (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3/) and visualized in PyMOL version 1.7.6.3 (https://pymol.org/).

In 2012, a database of tumor homing peptide (THP) namely TumorHoPe developed by Kapoor et al., stored comprehensive information of THPs to aid in designing peptide-based drugs and drug-delivery systems (Kapoor et al., 2012). To the best of our knowledge, there is only one computational model developed in this regard. (Sharma et al. (2013)) were the first to propose a sequence-based predictor called TumorHPD using support vector machine (SVM) as a prediction model and amino acid composition, dipeptide composition and binary profile patterns as feature input for discriminating THPs from non-THPs in peptide sequences. TumorHPD was able to yield a high accuracy of 86.56%. The aforementioned method has its own merit and played a major role in stimulating the development of this area although, bearing few limitations. Firstly, the peptide sequences in their datasets shared high-sequence similarity. It has been demonstrated that their datasets would lack statistical representations. Secondly, TumorHPD was trained and tested with such a dataset might yield overestimated prediction results. Lastly, the SVM model is well-recognized as a black-box approach that does not allow itself to interpretability, but the characterization and analysis of THPs are important for both basic research and drug development.

Given the aforementioned limitations of the existing method, it is desirable to develop an effective and interpretable a sequence-based model for discriminating THPs from non-THPs. Firstly, we constructed an objective and strict benchmark dataset by excluding peptides having a sequence identity of > 90%. Secondly, a prediction method named THPep was constructed by using an interpretable random forests (RF) method cooperated with amino acid composition, dipeptide composition and pseudo amino acid composition. This study utilizes the RF method because of its built-in ability of feature importance estimation. Thirdly, we estimated informative amino acids and dipeptides to provide insights of the biological implications of THPs. The prediction results over both 5-fold cross-validation and an independent dataset demonstrated that the proposed method THPep was superior to the existing method. Finally, a free web server was built to provide an efficient and useful tool for THP prediction.

2. Materials and methods

As elaborated by a series of publications from our group (Simeon et al., 2016, 2017; Pratiwi et al., 2017; Win et al., 2017; Shoombuatong et al., 2018a, b), to develop a useful and interpretable sequence-based predictor, the following five procedures should be considered: (i) establish or select a reliable benchmark dataset for training and validating the predictor and also calculating the dataset modelability as to evaluate the feasibility of obtaining a predictor with significant predictive power; (ii) represent protein or peptides sequences that can truly reflect their intrinsic properties to be predicted; (iii) introduce or develop an easily interpretable predictor that can provide a better understanding of the biological system; (iv) evaluate the predictive power of the predictor by performing rigorous cross-validation methods; (v) develop a user-friendly web-server that can easily reveal the desired results without needing to go through the mathematical and statistical details. Fig. 2 presents the workflow of the systems used to predict and analyze THPs, which involves dataset compilation, feature extraction, data splitting, model construction and evaluation and model post-analysis.

2.1. Benchmark datasets

A high quality benchmark dataset that was obtained from only experimentally verified peptides can guarantee the reliability and accuracy of a predictor. In this study, the benchmark datasets, i.e. Main (S_{Main}) and Small (S_{Small}) datasets, were derived from Sharma et al. (Sharma et al., 2013) to examine a predictor for its effectiveness in practical THP prediction as well as to fairly compare it with the existing method. The benchmark datasets S_{Main} and S_{Small} can be formulated as follows:

$$S_{Main} = S^{+}_{Main} \bigcup S^{-}_{Main} \quad (1)$$

$$S_{Small} = S^{+}_{Small} \bigcup S^{-}_{Small} \quad (2)$$

where S^{+}, S^{-} and the symbol $\bigcup$ represent the positive or THPs subset, the negative or Non-THPs subset and the union in the set theory, respectively. The subsets S^{+}_{Main} and S^{-}_{Main} contain 651 THPs and 651 Non-THPs, respectively, while the subsets S^{+}_{Small} and S^{-}_{Small} contain 469 THPs and 469 Non-THPs, respectively.

However, the peptides in Main (S_{Main}) and Small (S_{Small}) datasets share high-sequence similarity leading to a lack of statistical representation. As demonstrated in (Chou (2011)), a predictor trained and tested with such bias benchmark dataset, may yield misleading prediction results with overestimated accuracy. To remedy this problem in the proposed predictor, we constructed a new dataset, called Main90 dataset (S_{Main90}), by excluding peptides from the S_{Main} dataset with a sequence identity of 90% or above using CD-HIT (Huang et al., 2010). After such a procedure, the remaining dataset contained 176 THPs and 443 Non-THPs. The S_{Main90} dataset that can guarantee the reliability of the prediction model can be formulated as

$$S_{Main90} = S^{+}_{Main90} \bigcup S^{-}_{Main90} \quad (3)$$

where the subsets S^{+}_{Main90} and S^{-}_{Main90} contain 176 THPs and 443 Non-THPs, respectively. But, in the case of the S_{Small} dataset, we did not

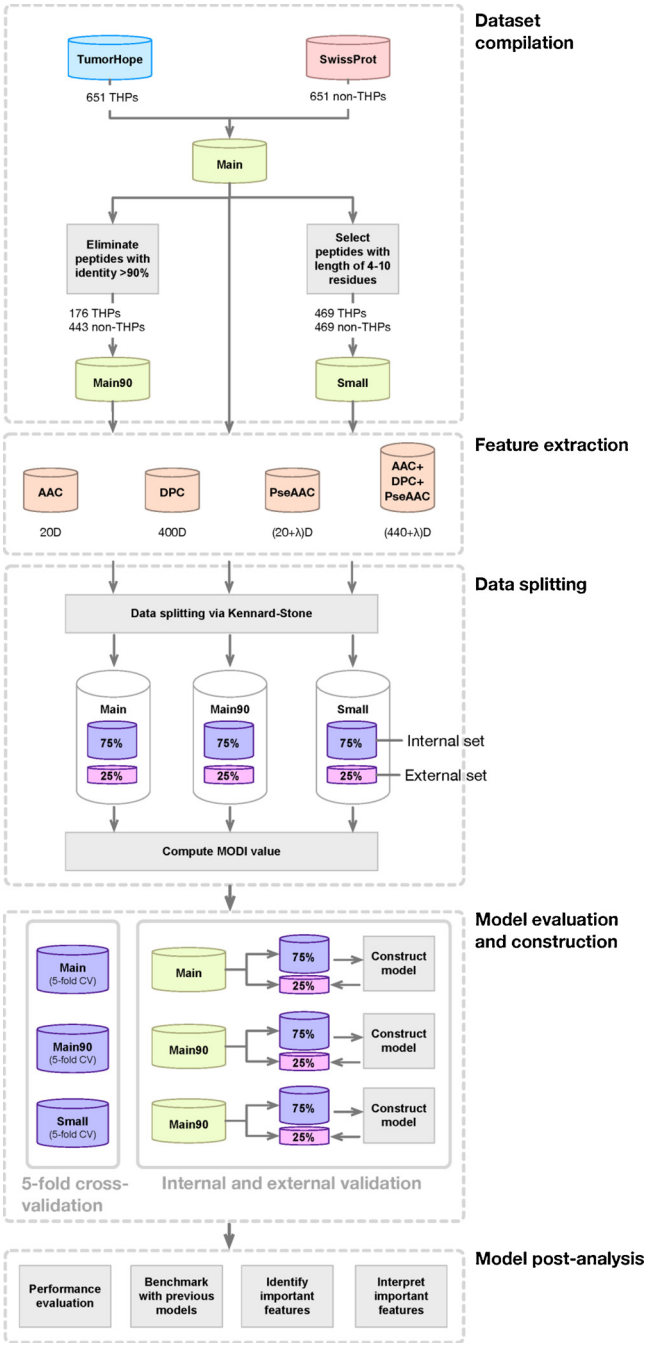

Fig. 2. Schematic illustration of the workflow for prediction and analysis of THPs.

Table 1
Summary of three benchmark datasets for evaluating the predictors of THPs. These benchmarked datasets were obtained from Sharma et al. (2013) (Sharma et al., 2013).

Dataset	Peptide identity	Whole dataset		Internal set		External set	
		THP	Non-THP	THP	Non-THP	THP	Non-THP
Main	100	651	651	490	490	161	161
Small	100	469	469	350	350	119	119
Main90	90	176	443	132	332	44	111

remove the homologous peptides because the short length of the peptides inside this dataset which may lead to a loss of information for establishing an efficient and effective predictor. Table 1 lists the distribution of THPs and non-THPs on all three benchmark datasets.

2.2. Feature representation

After obtaining the benchmark datasets, the next step is to represent each peptide into numerical vectors that can considerably depict the perspective of its biological and chemical properties. For peptide or protein sequences, the most classic and interpretable features are amino acid composition (AAC) and dipeptide composition (DPC). Until now, AAC and DPC have been applied in the investigation of many peptides and proteins, such as predicting HIV-1 CRF01-AE co-receptor usage (Shoombuatong et al., 2012), predicting protein crystallization (Shoombuatong et al., 2013; Charoenkwan et al., 2013), predicting the oligomeric states of fluorescent proteins (Simeon et al., 2016), predicting the bioactivity of host defense peptides (Simeon et al., 2017), predicting antifreeze proteins (Pratiwi et al., 2017), and predicting the hemolytic activity of peptides (Win et al., 2017).

According to the classical definition, the AAC, DPC and TPC features are expressed as a fixed length of 20, 400 and 8000 with each representing the frequency of one of 20, 400 and 8000 native amino acids, dipeptides and tripeptides, respectively, in the peptide sequence **P**. Thus, in the terms of AAC, DPC and TC features, a peptide **P** can be expressed by vectors with 20D and 400D (dimension) spaces, respectively as formulated by:

$$\mathbf{P} = [aa_1, aa_2, ..., aa_{20}]^T \quad (4)$$

$$\mathbf{P} = [dp_1, dp_2, ..., dp_{400}]^T \quad (5)$$

where T is the transpose operator, while aa_1, aa_2, …, aa_{20} and dp_1, dp_2, …, dp_{400} are occurrence frequencies of the 20 and 400 native amino acids and dipeptides, respectively, in a peptide sequence **P**.

As shown in Eq. 4 and 5, AAC and DPC features only provide the 20D AA and 400D DP compositions to represent a peptide **P**, but such three vectors defined by the concept of AAC and DPC may completely lose the sequence-order information. To deal with such dilemma, the pseudo amino acid composition (PseAAC) approach was proposed. According to Chou's PseAAC (Shoombuatong et al., 2018a), the general form of PseAAC for a peptide **P** is formulated by:

$$\mathbf{P} = [\Psi_1, \Psi_2, ..., \Psi_\Omega]^T \quad (6)$$

where the subscript Ω is an integer and its value as well as the components Ψ_1, Ψ_2, … will depend on the method used to extract the desired information from the amino acid sequence of a peptide **P**. Below, we describe the method used to generate the PseAAC feature from the three benchmark datasets to define the peptide sequences via Eq. 7.

2.3. Dataset modelability

One of the important factors for estimating the prediction performance of a QSAR model is the modelability of the dataset. In 2004, Golbraikh et al. introduced (Golbraikh et al., 2014; Fourches et al., 2016) the “Modelability Index” (MODI) as a simple and quick method to estimate the feasibility in developing robust and predictive QSAR model. The key idea behind this metric is that a pair of peptides having similar sequence patterns or characteristics are deemed to belong to the same biological class. In this study, the MODI score for each type of peptide features were calculated using an in-house developed R code. A threshold value of 0.65 separates the dataset into modelable or non-modelable classes. The steps involved in the calculation of the MODI score are described as follows:

Step 1: For a given dataset, the normalized Euclidean distance D'_{ij} between a pair of peptides M_i and M_j is computed as follows:

$$D_{ij} = \|\mathrm{M}_i - \mathrm{M}_j\| = \sqrt{\sum_{k=1}^{m} (M_{ik} - M_{jk})^2} \tag{7}$$

$$\bar{D}_i = \frac{\sum_{j=1}^{n} D_{ij}}{n-1} \tag{8}$$

$$D'_{ij} = \frac{D_{ij} - \min(\bar{D}_i)}{\max(\bar{D}_i) - \min(\bar{D}_i)} \tag{9}$$

where D_{ij}, $\bar{D}_i$ and n represents the distance scores between two peptides, the mean Euclidean distance and the number of peptides, respectively.

Step 2: For each peptide in a dataset, the MODI can be computed by identifying its first nearest neighbor (i.e. a peptide with the smallest Euclidean distance) belonging to the same or different class as follows:

$$MODI = \frac{1}{N_C} \sum_{l=1}^{N_C} \frac{N_i^{same}}{N_i^{total}} \tag{10}$$

where N_C is the number of classes (i.e. $N_C = 2$ denotes THP and non-THP), N_i^{same} is the number of peptides of the i^{th} class that have their first nearest neighbors belonging to the same i^{th} class and N_i^{total} is the number of peptides belonging to the i^{th} class. A dataset is considered to be modelable if the MODI index is greater than the threshold value of 0.65.

2.4. Random forest

(Breiman (2001)) proposed the random forest (RF) algorithm, which is an ensemble of tree-structured classifiers obtained by growing many weak classification and regression trees (CART) (Breiman et al., 1984) for enhancing the prediction performance. RF model takes advantage of two powerful machine-learning techniques: bagging (Breiman, 1996) and random feature selection. Growing many classification trees and using randomness in building RF model, yields a better prediction performance and enhances robustness of the model in a dataset. Each classification tree can be grown according to the following rules: (1) in bagging, each classification tree is trained on a bootstrap sample, which is a subset of the whole training N peptide. Peptides not included in the bootstrap sample are placed in the out-of-bag (OOB) set of M peptides, where $M \approx N/3$, which is then used to estimate the effectiveness of the RF model; (2) for each node of a classification tree, the RF model randomly selects *mtry* features and uses them to determine the best possible split with the Gini index as the splitting criteria; (3) each classification tree is grown to the largest extent possible i.e. no pruning is conducted. Herein, the RF classifier was established using the *randomForest* R package (Breiman, 2006). To enhance the performance of the RF model, two parameters namely *ntree* (i.e. the number of trees used for constructing the RF classifier) and *mtry* (i.e. the number of random candidate features) were subjected to optimization using the *caret* R package (Kuhn et al., 2017). The search for optimal values regarding the two parameters were carried out according to the following:

$$1 \leq mtry \leq 5\,,\ step\ \ \Delta = 1 \tag{11}$$

2.5. Feature importance

The RF method is useful for evaluating feature importance with the use of the OOB set. In order to provide a better understanding of the tumor homing activity of peptides, herein, the RF models were utilized for the identification of informative features for each type of peptide sequential feature. This is due to its efficient and effective built-in feature importance estimator. Generally, in the system of RF model construction, two measures based on the mean decrease of the Gini index (MDGI) and prediction accuracy are available for ranking feature importance. Considering the recommendations of (Calle and Urrea (2011)) that MDGI is more robust than the mean decrease of the prediction accuracy, we thus utilize the value of MDGI to rank the importance of each type of peptide sequential features. In the system of ranking feature importance, we constructed 10 RF models by fixing the *mtree* parameter to 100 and varying the *mtry* parameter setting from 2 to 20 ($mtry \in \{2,3,\ 5,7,\ 9,11,13,15,17,19,20\}$). The construction of multiple RF models for estimating important features can reveal potential bias of using a single value of *mtry* and thereby increases the reliability of the estimated informative features. Features with the largest MDGI value is considered to be the most important feature as it contributes the most to the prediction performance of the model.

2.6. Performance evaluation

In a statistical prediction, three cross-validation methods, namely sub-sampling test (2-, 5- or 10-fold cross-validation), jackknife test and independent (or external) testing dataset are often used to evaluate the prediction performance. Among the three aforementioned cross-validation methods, the jackknife test can yield a unique result for a given benchmark dataset, however, it is time-consuming. Thus, in this study, the 5-fold cross-validation (5-fold CV) and external testing dataset were used to evaluate the prediction performance of our models. The 5-fold CV approach was performed to benchmark and used to compared with the existing method (Sharma et al., 2013), while the external testing dataset was used to ensure that no overlap existed between the training and testing sets. Thus, it could be concluded that the predictor having high accuracy over the external testing dataset determines its high predictive ability on an unknown dataset. The procedure of external testing dataset is briefly described as follows. Firstly, the internal and external sets with the ratio of 3:1 were constructed using the Kennard-Stone algorithm via the R package *prospectr* (Stevens et al., 2015). Subsequently, the internal set was used to estimate the parameter of the model via a 5-fold CV and further, a prediction model with the optimum parameter was constructed to determine the class label on the external set.

In order to evaluate the prediction ability of the model, the following sets of four metrics are used as follows:

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{12}$$

$$Sn = \frac{TP}{(TP + FN)} \tag{13}$$

$$Sp = \frac{TN}{(TN + FP)} \tag{14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

where *Ac*, *Sn*, *Sp* and MCC are called accuracy, sensitivity, specificity and Matthews coefficient correlation, respectively. *TP*, *TN*, *FP*, and *FN* represent the instances of true positive, true negative, false positive and false negative, respectively. Moreover, in order to evaluate the prediction performance of models using threshold-independent parameters, the receiver operating characteristic (ROC) curves were plotted by the *pROC* package in the R software (Robin et al., 2011). The area under the ROC curve (auAUC) was used to measure the prediction performance, where AUC values of 0.5 and 1 are indicative of perfect and random models, respectively.

2.7. Physicochemical properties

When generating the PseAAC feature of a peptide, the following parameter set was considered: (1) the PseAAC mode; (2) the physicochemical properties (PCPs) used; (3) the two parameters (The lambda parameter λ and the weighting factor w, where the parameter λ represents the global or long-range sequence order effect, while the parameter w is the weight factor of a long-range effect usually within

the range of 0–1.

The parameter optimization procedure of PseAAC feature is described as follows. Firstly, the PseAAC model was set by using the default. Secondly, we selected the informative PCPs set from 531 PCPs (Kawashima et al., 2007) using the values of MDGI. During the determination of optimal subset of PCPs procedure, the RF model was constructed using the 5,10,15,20,… top-ranked informative PCPs and examined with 5-fold CV on the three benchmark datasets S_{Main}, S_{Small} and S_{Main90}. Finally, after obtaining the optimal PCPs subset, adjustments were made to the two parameters (λ and w) in order to achieve the best prediction accuracy. The search for the optimal values of the two parameters were carried out according to the following:

$$0.1 \le w \le 5\,,\ step\ \ \Delta = 0.1 \tag{16}$$

According to Eq. 16, total of 10 × 10 = 100 individual combinations should be investigated for determining the optimal parameter combination. Herein, the 5-fold CV approach was used to assess the performance for each parameter combination in the process of parameter optimization.

2.8. Reproducible research

To ensure the reproducibility of the models proposed herein, all R codes and the three benchmark datasets used in the construction of the predictive models, graphical figures and the THPep web server are available on GitHub at https://github.com/shoombuatong2527/thpep.

3. Results

3.1. Parameter determination

As mentioned above, the PseAAC feature depends on the PCPs and the two parameters (λ and w). At this step, the RF model was constructed using the 5, 10, 15, …, top-ranked informative PCPs and assessed by the 5-fold CV approach. The feature set with the highest prediction performance (Ac and MCC) were considered. Table S1 shows the performance comparisons among different subsets of informative PCPs on the three benchmark datasets. As noticed in Table S1, the 15, 30 and 25 top-ranked informative PCPs for S_{Main}, S_{Small} and S_{Main90} datasets showed the highest value of Ac/MCC of 84.56%/0.69, 81.77%/0.64 and 88.51%/0.71, respectively. Additional details of the informative PCPs on the three benchmark datasets are summarized in Table S2. After obtaining the optimal subset of informative PCPs, the two parameters of λ and w were estimated using the grid search within the range as mentioned in Eq. 16. The parameter combinations having the lowest prediction error was used to construct the THP predictor. In the present study, the optimal combination parameters λ and w for S_{Main}, S_{Small} and S_{Main90} datasets were 2/0.1, 3/0.6 and 1/0.5, respectively, in the following predictor development.

3.2. Prediction performance

In this study, we made an effort to develop a THP predictor that utilizes the RF model as the prediction model and AAC, PseAAC, DPC and all combination features, i.e. the combination of AAC + PseAAC, AAC + DPC, DPC + PseAAC and AAC + PseAAC + DPC, as the input features. In order to determine the features that are beneficial to the prediction of THPs, RF models cooperating with the different features were performed on the three-benchmark datasets via 5-fold CV. Moreover, in order to avoid the influence of redundancy, the proposed predictor was trained and tested with the S_{Main90} dataset having a sequence identity of 90% or less. Furthermore, in order to describe the predictable capability of our method on unknown samples, data splitting was performed for a 100 independent iterations on the three benchmarked datasets. The number of peptides was then used to create the models which were further evaluated using two cross-validation methods as shown in Table 1.

Table 2
Summary of MODI index as derived from various types of peptide features on both internal and external sets.

Feature	Internal set			External set		
	Main	Small	Main90	Main	Small	Main90
AAC	0.71	0.71	0.76	0.72	0.73	0.70
PseAAC	0.75	0.72	0.74	0.79	0.70	0.73
DPC	0.66	0.69	0.68	0.62	0.64	0.74
AAC + PseAAC	0.75	0.72	0.74	0.79	0.70	0.74
AAC + DPC	0.70	0.71	0.75	0.68	0.68	0.77
PseAAC + DPC	0.75	0.72	0.74	0.79	0.69	0.74
AAC + PseAAC + DPC	0.75	0.72	0.74	0.79	0.70	0.74

Before construction of the predictive model, the modelability of the dataset was assessed by computing the MODI index. This metric helps to estimate the feasibility of obtaining robust and reliable predictive model by using the three types of peptide features and their possible combinations for both internal and external sets. For binary classification problem, if the value of MODI index is greater than 0.65, the dataset is considered to be reliable for classification modelling. As seen in Table 2, most features from internal and external sets met these criteria with the exception of models built using DPC features on an external set.

Table 3–4 and Fig. 3 list the results of performance comparisons between different peptide features as evaluated by 5-fold CV and external testing dataset, respectively. Moreover, Figs. 3 shows the ROC curves of the two cross-validation methods. As noticed in Table 3, the combination feature of AAC + PseAAC affords the highest prediction performance in terms of Ac (86.10% and 90.80%) and MCC (0.72 and 0.77) on S_{Main} and S_{Main90} datasets, respectively. Meanwhile, the AAC feature performed well with the second highest Ac of 85.71% and 89.66% on S_{Main} and S_{Main90} datasets, respectively. In the case of the S_{Small} dataset, the AAC feature reached a maximum Ac of 83.37%, while the combination features of AAC + DPC and AAC + PseAAC performed well with the second and third highest Ac of 83.26% and 82.94%, respectively.

To infer the true predictive power of the proposed THP predictor, performance comparisons as evaluated by external testing dataset were also performed as shown in Table 4 and Fig. 3. Interestingly, the results of the performance comparisons were well correlated with the results derived from the 5-fold CV method (Table 3). As seen in Table 4, the combination feature of AAC + PseAAC and AAC + DPC reached a maximum Ac of 85.28%/83.74%, 90.13%/89.47% and 77.78%/80.77% on the S_{Main}, S_{Main90} and S_{Small} datasets, respectively. Furthermore, the combination feature of AAC + PseAAC outperformed the three single peptide features with an improvement of greater than 4–13% on MCC under the external testing dataset.

The results of performance comparisons can be briefly summarized as follows. Each of the three single peptide features can be effectively used to discriminate THPs from non-THPs. The AAC feature was observed as the most beneficial to be develop as a predictor than the DPC and PseAAC features. In the case of the combination features, the AAC + PseAAC is the effective and efficient peptide feature that performed well on the three benchmark datasets over the two standard cross-validation methods. All of the prediction results were seen to be well related with the MODI values. For convenience of the subsequence description, we will refer to this method as THPep.

3.3. Comparison with the existing prediction method

It is necessary to compare the proposed method with the existing method. To the best of our knowledge, there is only one other study which focused on the prediction of tumor homing peptides (Sharma

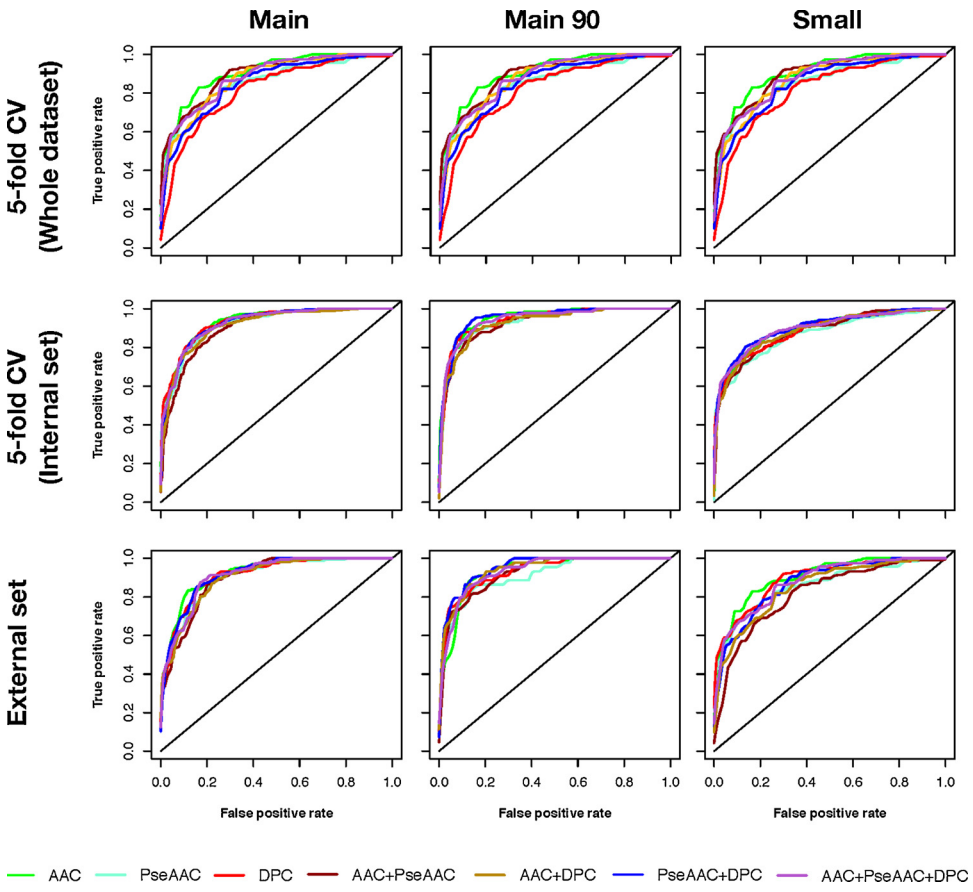

Fig. 3. ROC curve of RF models as assessed by 5-fold cross-validation. Models were evaluated via 5-fold CV on the whole dataset (fist row), 5-fold CV on the internal set (second row) and external test (third row). Performance of the three benchmarked datasets consisting of Main, Main 90 and Small are shown in the first, second and third columns, respectively.

Table 3
Performance comparison of RF models with various classes of protein sequences assesses by 5-fold cross-validation.

Dataset	Feature	Ac (%)	Sn (%)	Sp (%)	MCC	auAUC
Main	AAC	85.71	87.20	84.34	0.71	0.93
	PseAAC	84.56	84.72	84.40	0.69	0.92
	DPC	82.33	80.82	84.01	0.65	0.91
	AAC + PseAAC	86.10	87.07	85.18	0.72	0.93
	AAC + DPC	84.41	85.67	83.23	0.69	0.92
	PseAAC + DPC	84.72	84.98	84.45	0.69	0.92
	AAC + PseAAC + DPC	84.33	85.09	83.61	0.69	0.93
Small	AAC	83.37	81.24	85.81	0.67	0.91
	PseAAC	81.77	78.88	85.31	0.64	0.89
	DPC	80.06	80.00	80.13	0.60	0.87
	AAC + PseAAC	82.94	80.00	86.52	0.66	0.91
	AAC + DPC	83.26	80.95	85.94	0.67	0.91
	PseAAC + DPC	82.52	78.72	87.47	0.66	0.89
	AAC + PseAAC + DPC	82.62	79.54	86.43	0.66	0.90
Main90	AAC	89.66	90.22	87.92	0.74	0.95
	PseAAC	88.51	90.24	83.54	0.71	0.93
	DPC	86.86	87.79	83.80	0.67	0.92
	AAC + PseAAC	90.80	91.80	87.97	0.77	0.95
	AAC + DPC	88.51	88.05	90.15	0.71	0.95
	PseAAC + DPC	87.19	86.15	91.53	0.68	0.94
	AAC + PseAAC + DPC	88.34	88.03	89.47	0.71	0.95

et al., 2013). Thus, the prediction results between THPep and the existing method (TumorHPD (Sharma et al., 2013)) was compared. In 2013, TumorHPD was developed using a computational model based on the SVM method cooperating with AAC feature on the NTCT10 dataset. For facilitating comparison, the same benchmark datasets (S_{Main} and S_{Small}) and cross-validation method (5-fold CV) were employed. The results of the performance comparison between THPep and TumorHPD are listed in Table 5.

The reported results of TumorHPD as shown in Table 5 come directly from (Sharma et al., 2013). Since we are the first to perform a prediction model using a non-redundant dataset S_{Main90}, Table 5 does not show the prediction result of TumorHPD on this dataset. As noticed in Table 5, THPep outperformed TumorHPD with improvements of > 6-8%, > 2% and 2–3% for Sn, MCC and auAUC, respectively, as evaluated by the 5-fold CV approach. Although the Ac and Sp obtained from THPep were little lower than TumorHPD as evaluated by 5-fold CV approach, the MCC and auAUC are the most reliable statistic measures and are less influenced by an imbalanced dataset (Ganguly and Sadaoui, 2017). Considering that the external testing dataset is the most rigorous cross-validation method and a high value of MCC and auAUC are preferable, it could be stated that THPep is not only more effective than TumorHPD, but also more stable.

To further verify the power of THPep, some conventional classifiers such as *k*-nearest neighbor (*k*-NN), decision tree (DT), artificial neural network (ANN) and support vector machine (SVM) were implemented for performance comparisons. Such classifiers were tested on the three benchmark datasets and implemented using R programming (Gentleman, 2008). Table 6 lists the performance comparisons of THPep with the conventional classifiers based on the combination features of AAC and PseAAC. By observing the results listed in Table 6, we can clearly demonstrate that THPep outperforms all the four classifiers by achieving the highest Ac and MCC of 90.80 and 0.77, respectively as evaluated by 5-fold CV. Moreover, these results also indicated that the combination feature of AAC and PseAAC (83.79% Ac) are more effective than the proposed features in the work of (Sharma et al. (2013)) (82.52% Ac). The comparison results above leave no doubt that, THPep, as proposed in this study is quite promising and holds potential to become a useful tool for discriminating THPs from non-THPs. Furthermore, the proposed method, THPep, can play a complementary role to the existing method in the same area.

Table 4
Performance comparison of RF models with various types of protein sequences assessed by the 5-fold cross-validation and external sets.

Dataset	Feature	5-fold CV test					External testing dataset				
		Ac (%)	Sn (%)	Sp (%)	MCC	auAUC	Ac (%)	Sn (%)	Sp (%)	MCC	auAUC
Main	AAC	84.53	85.47	83.63	0.69	0.92	83.13	82.21	84.05	0.66	0.91
	PseAAC	83.50	82.63	84.42	0.67	0.92	83.44	82.82	84.05	0.67	0.91
	DPC	80.84	77.51	85.08	0.62	0.91	78.83	82.82	74.85	0.58	0.89
	AAC + PseAAC	84.63	86.11	83.27	0.69	0.93	85.28	82.82	87.73	0.71	0.92
	AAC + DPC	85.25	85.68	84.82	0.70	0.92	83.74	78.53	88.96	0.68	0.91
	PseAAC + DPC	83.81	83.13	84.52	0.68	0.92	82.52	82.21	82.82	0.65	0.90
	AAC + PseAAC + DPC	84.84	85.86	83.86	0.70	0.92	84.05	79.75	88.34	0.68	0.92
Small	AAC	81.68	80.55	82.89	0.63	0.91	78.63	83.76	73.50	0.58	0.87
	PseAAC	78.27	75.84	81.19	0.57	0.87	74.36	76.07	72.65	0.49	0.82
	DPC	79.69	78.17	81.38	0.59	0.87	72.22	75.21	69.23	0.45	0.82
	AAC + PseAAC	80.68	78.88	82.73	0.61	0.89	77.78	85.47	70.09	0.56	0.85
	AAC + DPC	80.40	78.16	83.02	0.61	0.90	80.77	83.76	77.78	0.62	0.87
	PseAAC + DPC	79.69	75.30	85.91	0.60	0.88	77.35	82.05	72.65	0.55	0.84
	AAC + PseAAC + DPC	80.11	77.60	83.12	0.60	0.89	78.21	81.20	75.21	0.57	0.85
Main90	AAC	87.53	88.51	84.40	0.69	0.94	88.82	93.52	77.27	0.72	0.95
	PseAAC	87.96	89.94	82.35	0.70	0.93	84.87	88.89	75.00	0.63	0.91
	DPC	85.12	85.79	82.65	0.62	0.93	88.16	90.74	81.82	0.72	0.93
	AAC + PseAAC	89.72	90.88	86.32	0.74	0.94	90.13	94.44	79.55	0.76	0.93
	AAC + DPC	85.34	85.44	84.95	0.63	0.94	89.47	97.22	70.45	0.74	0.94
	PseAAC + DPC	84.46	83.96	86.75	0.60	0.93	88.16	96.30	68.18	0.70	0.94
	AAC + PseAAC + DPC	85.78	84.57	91.36	0.64	0.95	89.47	97.22	70.45	0.74	0.95

Table 5
Performance comparisons between THPep and TumorHPD over 5-fold cross-validation.

Dataset	Method	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
Main	TumorHPD	86.56	80.63	89.71	0.70	0.91
	THPep	86.10	87.07	85.18	0.72	0.93
Small	TumorHPD	81.88	73.13	90.92	0.65	0.88
	THPep	83.37	81.24	85.81	0.67	0.91
Main90	TumorHPD	–	–	–	–	–
	THPep	90.80	91.8	87.97	0.77	0.95

Table 6
Performance comparisons between THPep and some conventional methods over 5-fold cross-validation.

Dataset	Method	Ac (%)	Sn (%)	Sp (%)	MCC
Main	*k*-NN	79.95	75.26	86.79	0.61
	DT	80.57	80.24	80.90	0.61
	ANN	81.87	81.30	82.47	0.64
	SVM	83.79	83.33	84.27	0.68
	THPep	86.10	87.07	85.18	0.72
Small	*k*-NN	79.00	82.54	76.15	0.58
	DT	76.97	77.56	76.41	0.54
	ANN	77.08	79.53	75.00	0.54
	SVM	79.10	81.52	77.03	0.58
	THPep	83.37	81.24	85.81	0.67
Main90	*k*-NN	81.94	67.93	88.00	0.57
	DT	84.40	73.68	88.58	0.62
	ANN	85.88	77.11	89.16	0.65
	SVM	86.54	85.07	86.95	0.66
	THPep	90.80	91.8	87.97	0.77

3.4. Mechanistic interpretation of feature importance

To the best of our knowledge, this study represents the first systematic effort for the selection and characterization of relevant features for THPs prediction. The characterization and analysis of feature importance can provide a unique understanding of THPs. The ability of THPs to recognize and bind with tumor cell receptors and tumor vasculature is considered their most important attribute. In order to function as a homing peptide, the presence of certain amino acid residues are required. In that regard, the existence of certain key residues in the sequence of THPs and their associated functions that determine activity will be discussed herein. Moreover, in this study, simple and easily interpretable features consisting of AAC and DPC were used. The efficient built-in feature importance estimator of the RF method was utilized to determine informative features for AAC and DPC. The value of MDGI was used to rank feature importance. The most important features are represented with the largest value of MDGI. The feature importance for AAC and DPC is shown in Fig. 4. The top three informative amino acids were shown to be Cysteine (Cys or C), Tryptophan (Trp or W) and Arginine (Arg or R) having MDGI values of 139.48, 46.56 and 45.40, respectively, while RC, GR, CR, and CG were the top four informative dipeptides having MDGI values of 9.22, 8.74, 8.41 and 7.86, respectively. Therefore, it can be inferred that peptides containing such amino acids and dipeptides have a higher chance of possessing THP activity. Interestingly, our results are consistent with the observations made by Sharma et al., where C, R, and W, were among the most abundant residues in THPs and are considered as the preferred residues present at the first position of a peptide sequence (Sharma et al., 2013).

3.5. THPep web server

To enable application of the proposed method, a user-friendly and freely accessible web server called THPep (i.e. based on the S_{Main} dataset) was created using the Shiny package under the R programming environment. As mentioned above, the best prediction performance was afforded by the RF model that makes use of the combination of AAC + PseAAC features. However, in practice, the length of this combination feature is a 420 + λ-dimensional vector, thereby resulting in a time-consuming procedure for both the construction of the prediction model and the prediction of the desired peptide sequence. Thus, selected AAC feature with 20-dimensional vector, which were found to be comparable with the aforementioned feature, was used for the construction of the THPep web server. Screenshots of the THPep web server is shown in Fig. 5. Furthermore, a step-by-step guide on how to use the THPep web server is given as follows:

Step 1. Browse to the THPep web server at http://codes.bio/THPep/ and you will see the top page of THPep on your computer screen as shown in Fig. 5(A).

Step 2. Either type or copy/paste the query peptide sequence into the Input box or upload the file of peptide sequences by clicking on the

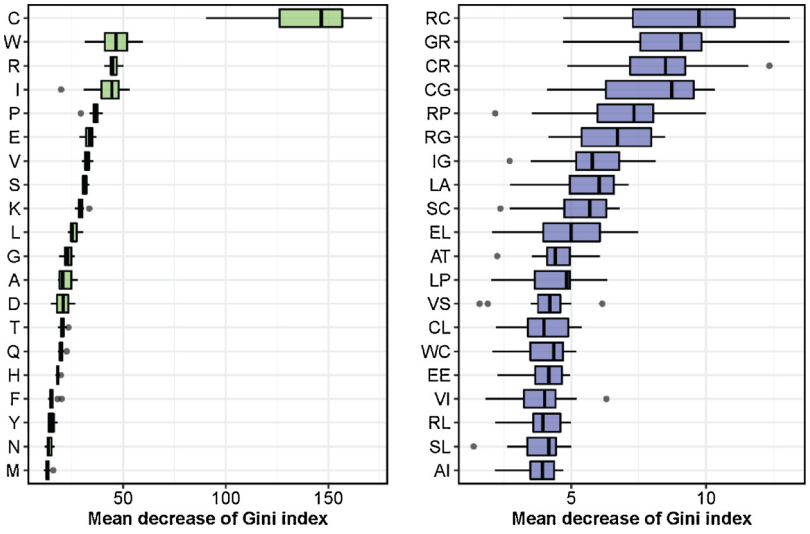

Fig. 4. Feature importance as deduced from the mean decrease of Gini index. Amino acid and dipeptide compositions are shown in the left and right panels, respectively.

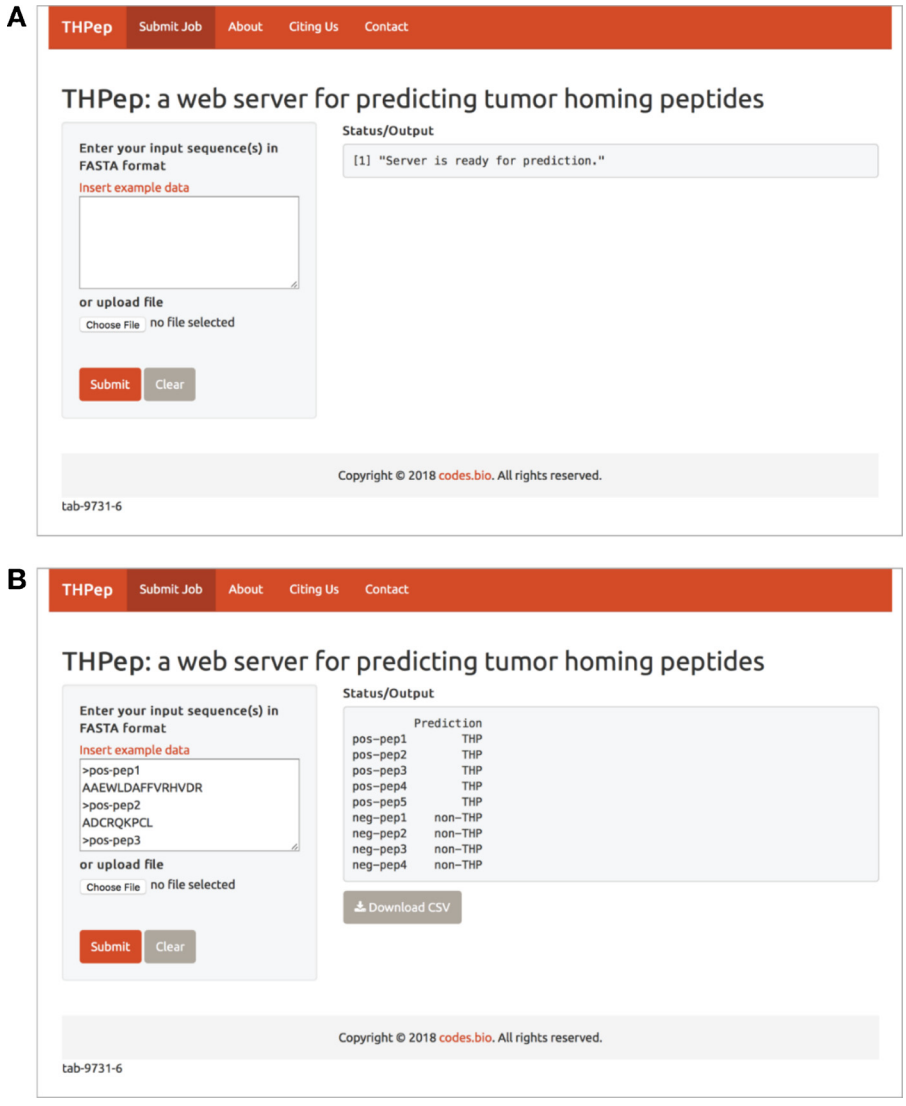

Fig. 5. Screenshot of the THPep web server. Before (A) and after (B) the initiation of the prediction process is shown.

Choose file button. The input peptide sequence should be in the FASTA format. Finally, press on the *Submit* button to initiate the prediction process.

Step 3. Click on the *Submit* button to initiate the modeling process and retrieve the prediction results, as shown in Fig. 5(B). Users can also download the prediction results as a CSV file by pressing on the *Download CSV* button.

Additionally, to maximize convenience, users could also build the THPep web server on their own computer by using the following code in an R environment:

shiny::runGitHub('THPep', 'chaninlab', subdir = "THPep_shiny _server")

Before running the aforementioned code, users must install prerequisite R packages by using the following code:

install.packages(c('shiny', 'shinyjs', 'shinythemes', 'protr', 'seqinr', 'randomForest', '
markdown'))

4. Discussion

The capability of current cancer therapies are limited by their ability in discriminating between tumor and healthy tissues. To solve such limitations, THPs have been developed for navigating therapeutic agents towards specific targets i.e. tumour tissues. Since the identification and development of new THPs using conventional experimental methods are labor intensive and time-consuming, computational models are the need of the hour for helping experimental scientists by offering analysis results. In the past decades, most model developers utilize complex computational models such as SVM, to predict many proteins and peptides, however those prediction models lacked mechanism prediction and interpretation. Rather than only focusing on the improvement of prediction results, model developers should also be concerned with the ability of their models which can help users in understanding and rationalizing their data of interest. In 2012, Huang et al. introduced a scoring card method (SCM) to predict protein solubility (Huang et al., 2012). For the prediction system, the SCM model used only a single threshold value to discriminate between soluble and insoluble proteins. In this work, they mentioned that an interpretable model with a comparable accuracy was more applicable. Most recently, Shoombuatong et al. provoked the importance and privilege of using interpretable models for researchers and scientists (Shoombuatong et al., 2017). Therefore, the goal of this study is to develop a systematic effort for predicting and analyzing THPs.

Table 3–4 and Fig. 3 compare the quality of prediction performances of RF model using seven feature sets as assessed by 5-fold CV and external testing dataset: AAC, PseAAC, DPC, AAC + PseAAC, AAC + DPC, PseAAC + DP and AAC + PseAAC + DPC. The performance comparisons on S_{Main}, S_{Small} and S_{Main90} datasets showed that the combination set of AAC + PseAAC resulted in superior predictions. Our prediction results was also shown to be consistent with previous related studies (Pratiwi et al., 2017; Win et al., 2017; Ganguly and Sadaoui, 2017). For instance, (Liaw et al. (2013)) developed an antibody amyloidogenesis model by using RF cooperating with AAC, DPC and PCP. This study suggested that the combined usage of the three feature types might improve prediction performances due to the fact that the usage of multiple features can capture the pattern of dataset more than by using a single feature. From the prediction results on each of the three basic features, the simple AAC feature performed better than PseAAC and DPC and also provided comparable results with the optimal combination sets mentioned above. Previously, the AAC feature has been widely used in the prediction and analysis of various protein and peptide functions, such as predicting the hemolytic activity of peptides (Win et al., 2017) or antifreeze proteins (Pratiwi et al., 2017).

As aforementioned, the top three informative amino acids were shown to be Cysteine (Cys or C), Tryptophan (Trp or W) and Arginine (Arg or R). Owing to its unique ability to form disulfide bonds through sulfhydryl side chains, Cys plays an important role in the function of THPs due to their abundance in the sequences. It has been reported that the cyclic peptide with two disulfide bonds, such as CDCRGDCFC, selectively accumulates in α_v integrins receptors in tumor blood vessels (Pasqualini et al., 1997). Moreover, another cyclic homing peptide i.e. cCPGPEGAGC (PEGA) containing one disulfide bridge is known to concentrate in the breast tumor vasculature and as a result, increases the efficacy of its conjugated-drug (Myrberg et al., 2008). Furthermore, a major limitation to the therapeutic effects of THPs is their short half-life in circulation, caused by renal clearance. Hence, coupling the peptide with a carrier molecule provides an effective strategy for prolonging its half-life by limiting renal clearance. This effect was demonstated by Pang et al. whereby the authors added a free Cys residue in the cyclic iRGD tumor-targeting peptide (CRGDK/RGPD/EC) that provided an extension to the half-life of the peptide and improved its accumulation in tumors. The added Cys conjugated with serum albumin through a disulfide bond and thus, afforded a higher molecular weight to the peptide that was above the renal clearance limit (Pang et al., 2014).

Furthermore, Colombo et al. investigated the importance of Cys residues using both, cyclic (CNGRC) and linear (GNGRG) peptides containing the NGR (Asn-Gly-Arg) motif using molecular dynamic simulations. The primary features under investigation were the stabilization of peptide conformations and their ability to deliver anti-tumor compounds (i.e. TNF-α) to tumor vessels. However, the authors found that the linear peptide containing Gly residues (GNGRG) had the propensity to form β-turn thus, lowering tumor targeting efficiency (Colombo et al., 2002) (Huang et al., 2012). The propensity of the linear peptide to form turns may favour the NGR motif through intramolecular stabilizing interactions (i.e. hydrogen bonding) thus, allowing them to be recognized by receptors and mediate anti-tumor compounds to tumor cells, albeit with lower efficiency. This in turn highlights the critical importance of Cys disulfide bridge formation in the enhanced stabilization and increased tumor targeting efficieny. Cyclic CNGRC peptides are known to bind with aminopeptidase N receptors expressed on the surface of cancer cells. Cys residues on the peptide formed a disulfide bridge which allowed the exposure of the N-terminus to bind with the active site of receptors, while the C-terminus carries the cargo (drug or chemical tag) (Graziadio et al., 2016).

Trp represents an amino acid having a highly hydrophobic indole ring side chain. Trp is important for the activity of THPs mainly due to the fact that their receptors reside in the membrane of tumor cells. Although considered a rare aromatic residue, Trp is highly enriched in membrane proteins (Von Heijne, 2007). A molecular dynamic study revealed the important role of Trp as a response modulator of hydrophobic mismatch to stabilize protein membrane anchoring in the phospholipid membrane (Jesus and Allen, 2013). It is likely that THPs containing Trp will be able to attach in the membrane protein of tumor cells through similar mechanisms as membrane protein anchoring. In that regard, an engineered M13 filamentous phage used for screening proteins present in the cell cytoplasm known as iPhage, was recently developed. The CPP sequence (RQIKIWFQNRRMKWKK) was attached to the pVIII capsid of the phage to allow internalization of the iPhage. Intriguingly, total functional loss of the iPhage was observed with Ala residue replacements, suggesting that the two Trp residues of the peptide sequence are essential for the intracellular translocation (Rangel et al., 2013).

In addition, the presence of Arg in a peptide sequence with the unique ability of conducting tumor homing, has been well demonstrated in both the NGR motif discussed above as well as in the RGD (Arg-Gly-Asp) motif (Pasqualini et al., 1997). These two motifs represent the first-generation of tumor homing peptides that appear to be independent of tumor type, indicating that their receptors are unregulated during angiogenesis. The RGD peptides have a high affinity towards integrins $\alpha_v\beta_3$ and $\alpha_v\beta_5$ which are specifically over-expressed during angiogenesis and nearly absent in the normal tissue (Varner and

Cheresh, 1996). One of the problems associated with drug permeation into tumor tissue is that the blood flow through tumor vessels may be abnormal (Hambley and Hait, 2009). Furthermore, it is well known that Arg is primarily involved in the improvement of blood circulation in the body through its metabolism of nitric oxide (Scibior and Czeczot, 2004). Thus, it can be inferred that peptides containing Arg as part of its tumor homing motif, may benefit from this essential amino acid that has also been demonstrated to protect against the early stages of cancer (Geiger et al., 2016). In addition, novel RGD peptides have been adapted to include a tissue penetrating motif that binds to integrins and then upon proteolytic cleavage, enables binding to neuropilin-1 (NRP-1), thus triggering tissue penetration (Sugahara et al., 2009, 2010). Besides, various forms of RGD (e.g. RGD-4C, iRGD and cRGD) have been used to target arthritis and various cancers by coupling with apoptotic peptides and anti-cancer drugs (Sugahara et al., 2010; Elayadi et al., 2007; Gerlag et al., 2001). Furthermore, the first cell penetrating homing peptide to enter clinical trials, Cilengitide (Merck) is a salt of cRGD that acts as an anti-angiogenic agent and $\alpha_v\beta_3$ and $\alpha_v\beta_5$ integrin antagonist (Eskens et al., 2003; Fink et al., 2010; Gilbert et al., 2011).

Furthermore, (Lee et al. (2013)) identified peptides using bioinformatics that has recently progressed to the pre-clinical phase whereby the basement membrane of endothelial cells were explored via protein-protein interactions. Similarly, the peptide-peptide combination of AARP (Ala-Ala-Arg-Pro), coupled with an assembly of the metalloproteinase inhibitory peptides (MMP-2 and 9; CTTHWGFTLC) were conjugated to two different anti-angiogenic agents. Thus, the recombinant AARP-CTT comprises one single compound which was designed to block angiogenesis and inhibit MMP2 and -9 at the same time. This conjugate has been used to target various human solid tumors in mouse models and shown to inhibit angiogenesis, tumor growth, and metastasis, efficiently (Wang et al., 2014). Additionally, a novel, nine amino acid long peptide (CGLSGLGVA) known as CooP, for use against malignant brain tumour as well as in drug delivery, has been ascertained. CooP binds to the over-expressed cancer cell only, mammary-derived growth inhibitor and has further demonstrated exceptional homing peptide capabilities with its targeting of tumor-associated blood vessels in the brain (Hyvonen et al., 2014).

Due to their high specificity, low antigenicity, flexibility, and simple production, peptides represent a promising new avenue of possibilities for the development of anti-cancer therapeutic agents for various tumors. In the long term, such an approach can become personalized as the current use of THPs for drug delivery, such that novel mechanisms can be targeted which cannot otherwise be reached with traditional small molecules. However, while the identification of specific THPs is relatively fast, the identification of their target receptors represents a bottleneck of the system. As such, no universal approach exists for drug design although promising methods are reliant on the binding affinity between peptide and receptor. Since THPs are administered with the aim that they bind to their targets via circulation, efficient, reliable and generalize attachment needs to be developed.

5. Conclusions

Much evidence has been acquired to indicate the favourable nature of tumor homing peptides as promising candidates for cancer therapy. Thus, accurate identification and efficient characterization based on a computational model of tumor homing peptides are essential for understating their role in drug development as well as reducing time and costs. In this study, an interpretable and powerful sequence-based model termed THPep has been developed by using an interpretable random forest classifier and multiple peptide features. Rigorous cross-validation using both 5-fold cross-validation and independent testing dataset have indicated that the proposed THPep model is extremely powerful and promising. It is anticipated that THPep might become an important tool for both basic research and drug development.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank TS Win and V Prachayasittikul for helpful comments and suggestion on this manuscript. This work is supported by the TRF Research Grant for New Scholar (No. MRG6180226) and the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund, the Office of Higher Education Commission and Mahidol University; the New Researcher Grant (A31/2561) from Mahidol University; and the Ph.D. scholarship from the Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

References

N. I. o. Health, 2017. Cancer Statistics of National Cancer Institute.
W. H. Organization, 2017. World Health Organization Cancer Fact Sheet.
Mäe, M., Myrberg, H., El-Andaloussi, S., Langel, Ü., 2009. Design of a tumor homing cell-penetrating peptide for drug delivery. Int. J. Pept. Res. Ther. 15, 11–15.
Myrberg, H., Zhang, L., Mäe, M., Langel, Ü., 2008. Design of a tumor-homing cell-penetrating peptide. Bioconjug. Chem. 19, 70–75.
Svensen, N., Walton, J.G.A., Bradley, M., 2012. Peptides for cell-selective drug delivery. Trends Pharmacol. Sci. 33, 186–192.
Laakkonen, P., Vuorinen, K., 2010. Homing peptides as targeted delivery vehicles. Integr. Biol. 2, 326–337.
Pasqualini, R., Koivunen, E., Rouslahti, E., 1997. Av integrins as receptors for tumor targeting by circulating ligands. Nat. Biotechnol. 15, 542–546.
Pasqualini, R., et al., 2000. Aminopeptidase N is a receptor for tumor-homing peptides and a target for inhibiting angiogenesis. Cancer Res. 60, 722–727.
Gautam, A., et al., 2014. Tumor homing peptides as molecular probes for cancer therapeutics, diagnostics and theranostics. Curr. Med. Chem. 21, 2367–2391.
Kapoor, P., Singh, H., Gautam, A., Chaudhary, K., Kumar, R., Raghava, G.P.S., 2012. TumorHoPe: a database of tumor homing peptides. PLoS One 7, 1–6.
Sharma, A., et al., 2013. Computational approach for designing tumor homing peptides. Sci. Rep. 3, 1607.
Simeon, S., et al., 2016. osFP: a web server for predicting the oligomeric states of fluorescent proteins. J. Cheminform. 8, 72.
Simeon, S., et al., 2017. PepBio: predicting the bioactivity of host defense peptides. RSC Adv. 7, 35119–35134.
Pratiwi, R., et al., 2017. CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. J. Chem. 2017.
Win, T.S., Malik, A.A., Prachayasittikul, V., Wikberg, J.E.S., Nantasenamat, C., Shoombuatong, W., 2017. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med. Chem. 9, 275–291.
Shoombuatong, W., Schaduangrat, N., Nantasenamat, C., 2018a. Towards understanding aromatase inhibitory activity via QSAR modeling. EXCLI J. 17, 688.
Shoombuatong, W., Schaduangrat, N., Nantasenamat, C., 2018b. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI J. 17, 734.
Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.
Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J., Samleerat, T., 2012. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. Comput. Biol. Med. 42, 885–889.
Shoombuatong, W., Huang, H.-L., Chaijaruwanich, J., Charoenkwan, P., Lee, H.-C., Ho, S.-Y., 2013. Predicting protein crystallization using a simple scoring card method. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) 23–30.
Charoenkwan, P., Shoombuatong, W., Lee, H.-C., Chaijaruwanich, J., Huang, H.-L., Ho, S.-Y., 2013. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PLoS One 8, e72368.
Golbraikh, A., Muratov, E., Fourches, D., Tropsha, A., 2014. Data set modelability by {QSAR}. J. Chem. Inf. Model. 54, 1–4.
Fourches, D., Muratov, E., Tropsha, A., 2016. Trust, but verify II: a practical guide to chemogenomics data curation. J. Chem. Inf. Model. 56, 1243–1252.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2006. random{F}orest: {B}reiman and {C}utler's random forests for classification and regression.
Kuhn, M., et al., 2017. caret: classification and regression training.
Calle, M.L., Urrea, V., 2011. Letter to the editor: stability of Random Forest importance measures. Brief. Bioinformatics 12, 86–89.

Stevens, A., Ramirez-Lopez, L., Stevens, M.A., Rcpp, L., 2015. prospectr. Miscellaneous Functions for Processing and Sample Selection of vis-NIR Diffuse Reflectance Data. pp. 32.
Robin, X., et al., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.
Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M., 2007. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36, D202–D205.
Ganguly, S., Sadaoui, S., 2017. Classification of imbalanced auction fraud data. Canadian Conference on Artificial Intelligence 84–89.
Gentleman, R., 2008. R Programming for Bioinformatics. Chapman and Hall/CRC.
Huang, H.-L., et al., 2012. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. BMC Bioinformatics 13, S3.
Shoombuatong, W., et al., 2017. Towards the revival of interpretable QSAR models. In: Roy, K. (Ed.), Advances in QSAR Modeling, pp. 3–55.
Liaw, C., Tung, C.-W., Ho, S.-Y., 2013. Prediction and analysis of antibody amyloidogenesis from sequences. PLoS One 8, e53235.
Pang, H.-b., et al., 2014. A free cysteine prolongs the half-life of a homing peptide and improves its tumor-penetrating activity. J. Control. Release 175, 48–53.
Colombo, G., et al., 2002. Structure-activity relationships of linear and cyclic peptides containing the NGR tumor-homing motif. J. Biol. Chem. 277, 47891–47897.
Graziadio, A., et al., 2016. NGR tumor-homing peptides: structural requirements for effective APN (CD13) targeting. Bioconjug. Chem. 27, 1332–1340.
Von Heijne, G., 2007. The membrane protein universe: What's out there and why bother? J. Intern. Med. 261, 543–557.
Jesus, A.J.D., Allen, T.W., 2013. The role of tryptophan side chains in membrane protein anchoring and hydrophobic mismatch. Biochimica et Biophysica Acta (BBA) - Biomembranes 1828, 864–876.
Rangel, R., et al., 2013. Targeting mammalian organelles with internalizing phage (iPhage) libraries. Nat. Protoc. 8, 1916–1939.
Varner, J.A., Cheresh, D.A., 1996. Integrins and cancer. Curr. Opin. Cell Biol. 8, 724–730.
Hambley, T.W., Hait, W.N., 2009. Is Anticancer Drug Development Heading in the Right Direction? Cancer Res. 69, 1259–1262.
Scibior, D., Czeczot, H., 2004. Arginine-metabolism and functions in the human organism. Postepy Hig. Med. Dosw. 58, 321–332.
Geiger, R., et al., 2016. L-arginine modulates t cell metabolism and enhances survival and anti-tumor activity. Cell 167, 829–842 e13.
Sugahara, K.N., et al., 2009. Tissue-penetrating delivery of compounds and nanoparticles into tumors. Cancer Cell 16, 510–520.
Sugahara, K.N., et al., 2010. Coadministration of a tumor-penetrating peptide enhances the efficacy of Cancer drugs. Science 328, 1031–1035.
Elayadi, A.N., et al., 2007. A peptide selected by biopanning identifies the integrin αvβ6 as a prognostic biomarker for nonsmall cell lung Cancer. Cancer Res. 67, 5889–5895.
Gerlag, D.M., et al., 2001. Suppression of murine collagen-induced arthritis by targeted apoptosis of synovial neovasculature. Arthritis Res. 3, 357.
Eskens, F.A.L.M., et al., 2003. Phase I and pharmacokinetic study of continuous twice weekly intravenous administration of Cilengitide (EMD 121974), a novel inhibitor of the integrins alphavbeta3 and alphavbeta5 in patients with advanced solid tumours. Eur. J. Cancer 39, 917–926.
Fink, K., et al., 2010. Long-term effects of cilengitide, a novel integrin inhibitor, in recurrent glioblastoma: A randomized phase {IIa} study. J. Clin. Oncol. 28, 2010.
Gilbert, M.R., et al., 2011. Cilengitide in patients with recurrent glioblastoma: the results of NABTC 03-02, a phase II trial with measures of treatment delivery. J. Neurooncol. 106, 147–153.
Lee, E., Koskimaki, J.E., Pandey, N.B., Popel, A.S., 2013. Inhibition of Lymphangiogenesis and angiogenesis in breast tumor xenografts and lymph nodes by a peptide derived from transmembrane protein 45A. Neoplasia 15, 112–IN6.
Wang, H., Yang, Z., Gu, J., 2014. Therapeutic targeting of angiogenesis with a recombinant CTT peptide-endostatin mimic-kringle 5 protein. Mol. Cancer Ther. 13, 2674–2687.
Hyvonen, M., et al., 2014. Novel Target for Peptide-Based Imaging and Treatment of Brain Tumors. Mol. Cancer Ther. 13, 996–1007.