Analytical Biochemistry 599 (2020) 113747

Contents lists available at ScienceDirect

Analytical Biochemistry

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

iTTCA-Hybrid: Improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation

Phasit Charoenkwan[a], Chanin Nantasenamat[b], Md Mehedi Hasan[c], Watshara Shoombuatong[b,*]

[a] Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, 50200, Thailand
[b] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, 10700, Thailand
[c] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan

ARTICLE INFO

Keywords:
Cancer immunotherapy
T-cell antigen
T-cell epitope
Random forest
Support vector machine
Machine learning

ABSTRACT

In spite of the repertoire of existing cancer therapies, the ongoing recurrence and new cases of cancer poses a challenging health concern that prompts for novel and effective treatment. Cancer immunotherapy represents a promising venue for treatment by harnessing the body's immune system to combat cancer. Therefore, the identification of tumor T cell antigen represents an exciting area to explore. Computational tools have been instrumental in the identification of tumor T cell antigens and it is highly desirable to attain highly accurate models in a timely fashion from large volumes of peptides generated in the post-genomic era. In this study, we present a reliable, accurate, unbiased and automated sequence-based predictor named iTTCA-Hybrid for identifying tumor T cell antigens. The iTTCA-Hybrid approach proposed herein employs two robust machine learning models (e.g. support vector machine and random forest) constructed using five feature encoding strategies (i.e. amino acid composition, dipeptide composition, pseudo amino acid composition, distribution of amino acid properties in sequences and physicochemical properties derived from the AAindex). Rigorous independent test indicated that the iTTCA-Hybrid approach achieved an accuracy and area under the curve of 73.60% and 0.783, respectively, which corresponds to 4% and 7% performance increase than those of existing methods thereby indicating the superiority of the proposed model. To the best of our knowledge, the iTTCA-Hybrid is the first free web server (Available at http://camt.pythonanywhere.com/iTTCA-Hybrid) for identifying tumor T cell antigens presented by the MHC class I. The proposed web server allows robust predictions to be made without the need to develop in-house prediction models.

1. Introduction

Cancer poses a major global health concern as it represents the leading causes of death by accounting for almost 10 million deaths in 2018 [1]. The dramatic increase in cancer prevalence over the past decades has called for intense efforts in prevention and treatment. Novel therapeutic agents for combating cancer is much sought after owing to the inherently high cancer recurrence despite the availability of an arsenal of therapeutic regimens for cancer treatment [2]. Small-molecule drugs against cancer have demonstrated potent therapeutic properties but the undesirable cytotoxic effects on healthy cells still pose a major impediment that calls for the search for potent yet safe therapeutic agents [3]. Immune therapy represents a promising therapeutic approach as it primes the body's natural defenses against cancer. Unlike deleterious side effects from chemo- and radiation therapies that are known to give rise to drug resistance, immune therapy has been shown to afford high selectivity and efficacy while causing lowered side effects [4]. As such, immunotherapy provides a new avenue of opportunities for the development of potent cancer treatment.

Cancer immunotherapy exert its activity via innate or adaptive immunity. The innate immunity serves as the first-line of defense and elicits non-specific response, for instance, via host defense peptides that target membranes of tumor cells owing to their inherent amphipathic properties [5,6]. On the other hand, the adaptive immunity acts via specific response mediated by B and T cells. In the case of T cells, T cell receptors (TCR) found at the surface encounters tumor antigens that are presented by the major histocompatibility complex (MHC) class I and II molecules found at the surface of antigen presenting cells (and hence its name) [7]. Tumors are immunogenic as they inherently harbor antigenic determinants (known as epitopes) as derived from peptides and proteins. A formidable challenge in the development of effective cancer

* Corresponding author.
E-mail address: watshara.sho@mahidol.ac.th (W. Shoombuatong).

https://doi.org/10.1016/j.ab.2020.113747
Received 16 March 2020; Received in revised form 13 April 2020; Accepted 16 April 2020
Available online 22 April 2020
0003-2697/

vaccines is the difficulty of epitopes that are capable of inducing a strong and selective immune response. Definitive evidence suggests that peptides derived from tumor-associated antigens can induce a peptide-specific T-cell immune response [8]. Furthermore, cancer neoantigens that are not present in normal cells represents a promising therapeutic opportunity as they represent specific targets and inherently leads to lower risk of autoimmunity and immune tolerance [9–11]. Taken together, the identification of T-cell epitopes in tumor antigens represent a lucrative first step towards the discovery and development of cancer immunotherapeutic peptides.

Although, experimental approaches are known as the most reliable way of characterizing the biological activities of T-cell epitopes in tumor antigens, it is time-consuming and costly. The availability of large volumes of peptides in several public databases, provides ample opportunity for the development of fast and intelligent computational models for identifying T-cell epitopes in tumor antigens, which is urgently needed for the development of cancer vaccines. Over the years, researchers have developed bioinformatics and immunoinformatics tool for facilitating peptide vaccine discovery [6,12–33]. Since epitopes play a crucial role for both clinical and basic science research, their robust characterization and analysis affords huge potential for vaccine design, disease prevention, diagnosis and treatment [34]. Thus far, there are four major approaches that have been employed for predicting epitopes namely sequence-based methods [35–37], structure-based methods [38], hybrid methods and consensus methods. For example, in 2004, Manoj Bhasin and Raghava [35] developed a sequence-based method known as the CTLPRED for predicting cytotoxic T lymphocyte (CTL) epitopes, which represent potential candidates in subunit vaccine design for various diseases by making use of antigenic sequence. The CTLPRED was developed using support vector machine (SVM) and artificial neural network (ANN) as well as quantitative matrix (QM). CTLPRED was performed on a non-redundant dataset of T cell epitopes and non-epitopes containing 1137 experimentally proven MHC class I restricted T cell epitopes. Prediction accuracies for QM-, ANN- and SVM-based methods were 70.0, 72.2 and 75.2%, respectively, as assessed by a leave-one-out cross-validation (LOOCV) test set. Finally, the machine learning methods were used for consensus and combined prediction of CTL epitopes. Finally, an online prediction server CTLPRED was implemented and freely accessible at http://www.imtech.res.in/raghava/ctlpred/. As mentioned above, the first important step for the discovery and development of cancer immunotherapeutic peptides is to identify T-cell epitopes in tumor antigens.

To the best of our knowledge, there is only one study focusing on the prediction of tumor T cell antigens represented in the MHC class I context, which is the TTAgP1.0 [39]. This method was developed via the use of the random forest (RF) model that makes use of two peptide feature types namely the relative frequency (Rfre) and amino acid composition (AAC) of peptides (PEP). Although, TTAgP1.0 has its own merit and reasonable prediction accuracies, there are four major limitations that needs to be addressed so as to improve the utilization of computational tools [39]. Firstly, Lissabet et al. [39] did not provide their dataset used. However, several researchers have mentioned the benefits of data sharing so as to facilitate reproducible research. Secondly, TTAgP1.0 was developed via the use of only two types of peptide features which might not be able to fully capture the significant information of tumor T cell antigens. Thirdly, TTAgP1.0 was implemented and performed on an imbalanced dataset (442 positive and 295 negative samples) that likewise contributes to classification errors and bias. Fourthly, no web server was provided from the study. Hence, its utility and usage is quite limited for the broad scientific community.

To address these aforementioned issues, the present study describes the development of a robust, unbiased and automated sequence-based predictor referred herein as the iTTCA-Hybrid for identifying tumor T cell antigens. Fig. 1 shows the workflow of the iTTCA-Hybrid for identifying tumor T cell antigens as represented in the MHC class I context. Since, previous studies did not publicly share the dataset used in the studies, therefore this study establishes the first publicly available benchmark dataset consisting of 529 tumor and 393 non-tumor T cell antigen so as to ensure the reproducibility of the proposed model. Secondly, this study employs two robust machine learning models (e.g. RF and SVM) in conjunction with five feature encoding strategies. Thirdly, to cope with the class imbalance problem and to improve the predictive power of the proposed model, extracted features were hybridized and the synthetic minority over-sampling technique (SMOTE) method was utilized to overcome possible bias arising from the class imbalance problem. On the basis of these comparative results, iTTCA-Hybrid can achieve significantly improved performance than those of existing methods thereby indicating the effectiveness and robustness of the proposed method. Finally, for the convenience of experimental biologists, the iTTCA-Hybrid web server was established and made freely available online at http://camt.pythonanywhere.com/iTTCA-Hybrid.

2. Materials and methods

2.1. Benchmark datasets

In practice, same training and independent sets should be employed for the development and assessment of the proposed model so as to provide a comprehensive and unbiased comparison [13,27–29,40,41]. However, previous studies [39] did not publicly share the dataset used in the study. In order to make a fair comparison, we generated an up-to-date benchmark dataset that contains the largest number of samples following the procedures described in the previous study [39] as follows. (i) Tumor T cell antigens were obtained from the TANTIGEN [42] and TANTIGEN 2.0 [43] whereby a total 727 MHC class I peptides was collected and considered as positive samples, (ii) Non-tumor T cell antigens were obtained from well-known IEDB database (https://www.iedb.org/). In order to construct the dataset of non-tumor T cell antigens, only T cell antigens that are reported to have no association to any disease were collected and considered as negative samples (containing 367 samples). (iii) Duplicate peptide sequences were removed. After performing such pre-processing, all peptides were unique and constituted the benchmark dataset consisting of 529 positive and 393 negative samples.

In order to examine the efficacy of the proposed model, an independent dataset (named TTCA-IND) was constructed by randomly selecting 20% from the benchmark dataset while treating the remaining samples as the training dataset (named TTCA-CV). Finally, the TTCA-CV dataset consisted of 470 positive and 318 negative samples while the TTCA-IND dataset consisted of 122 positive and 75 negative samples as summarized in Table 1. In order to ensure the reproducibility of the model proposed herein, datasets used in the construction of predictive models are shared publicly and is available on GitHub at https://github.com/Shoombuatong2527/Benchmark-datasets.

2.2. Feature representation

Given a peptide sequence (P) that can be represented as:

$$P = p_1 p_2 p_3 \cdots p_N \quad (1)$$

where p_i and N denote the ith residue in the peptide P and the peptide length, respectively. Note that residue p_i is in the set of natural amino acid, i.e. A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y.

The simplest feature encoding strategies used to represent a peptide sequence are AAC and dipeptide composition (DPC) that are expressed as fixed lengths of 20 and 400, respectively. Thus, a peptide P can be expressed by vectors with 20D and 400D (dimension) spaces for AAC and DPC features, respectively, as formulated by:

$$P = [aa_1, aa_2, ..., aa_{20}]^T \quad (2)$$

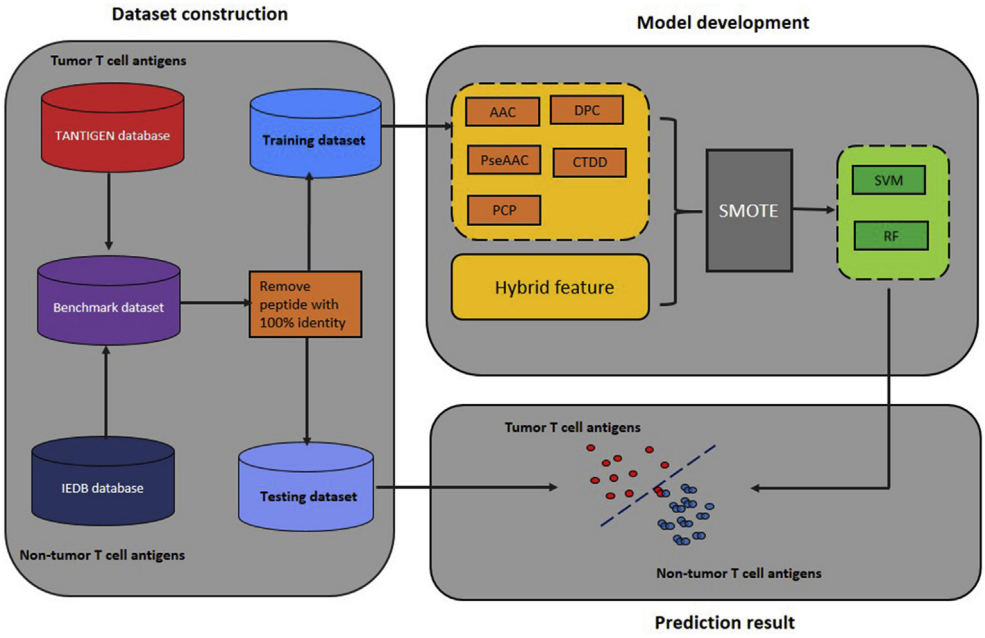

Fig. 1. Schematic framework of iTTCA-Hybrid for predicting tumor T cell antigens.

Table 1
Dataset distribution of peptides used in this study.

Dataset	Original	Non-duplicate peptides	Benchmark dataset	
			Positive	Negative
TTCA-CV	727	592	470	122
TTCA-IND	394	393	318	75

$$P = [dp_1, dp_2, ..., dp_{400}]^T \quad (3)$$

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

As for physicochemical (PCP) features, there are a total of 544 PCPs for amino acids as derived from version 9.0 of the Amino acid index database (AAindex) [44]. In this study, PCPs having NA values were excluded. Finally, a set of 531 PCPs (531D) were used to extract peptide sequences for the development of the model. Previously, PCPs are one of the most intuitive features associated with biophysical and biochemical reactions and is also referred to as easy and interpretable features [6,17,18,31,40,45].

To address the sequence-order information, the feature encoding strategy named pseudo amino acid composition (PseAAC) was proposed by Chou [46,47]. According to Chou's PseAAC, the general form of PseAAC for a peptide P is formulated by:

$$P = [\Psi_1, \Psi_2, ..., \Psi_u, ..., \Psi_\Omega]^T \quad (4)$$

where the subscript Ω is an integer to reflect the feature's dimension. The value of Ω and the component of Ψ_u, where u= 1,2, ...,Ω is dependent on the protein or peptide sequences.

Distribution of amino acid properties in sequences (CTDD) are often used to describe the overall composition of the amino acid properties of peptide sequences by using CTD descriptor having three types of descriptors, Composition (C), Transition (T), and Distribution (D) [48]. The CTDD feature is the distribution descriptor that describes the distribution of the first, 25%, 50%, 75%, and 100% amino acids of a particular property within the peptide sequence. Additional details of CTDD feature can be found in Refs. [48]. Previously, several studies have reported that AAC, DPC, PseAAC, PCP and CTDD features are effective features for predicting and analyzing various types of protein and peptide functions [6,13,18,30,33,40,41,49,50]. Thus, in this study, these five different types of sequence features were generated by using the iFeature module in the Python environment [48].

2.3. Synthetic minority over-sampling technique

From the point of view of machine learning, the class imbalanced dataset has a tendency to cause a prediction model to overfit as well as to perform poorly on the minority class. To address this problem, there are two widely used approaches containing undersampling and oversampling methods. The idea of undersampling methods is to create a new subset of the imbalanced dataset by eliminating some of the samples from the majority class, while oversampling methods create a new subset of the imbalanced dataset by creating new ones from the original minority class samples [51]. In this study, we employ the powerful SMOTE algorithm [52] for performing oversampling of the minority class (i.e., negative samples) by taking each minority class sample and introducing synthetic samples as summarized in Table 1.

2.4. Classifier selection

RF models were constructed according to the originally described algorithm [53]. RF is an ensemble-based ML algorithm capable of dealing with binary and multi-class classification problems and has been widely used in various biological problems [6,12,16–18]. To improve the prediction performance of classification and regression tree (CART) [53,54], These models were established by generating a number of CART classifiers. In RF, the out-of-bag (OOB) approach is employed for assessing the feature importance as follows: (i) two-thirds of the training data are used to construct the classifier while the remaining was used for evaluating the performance of the constructed classifier and (ii) the feature importance of each feature can be assessed by measuring the decrease in the prediction performance. It should be noted that the performance evaluation of the model can be either accuracy or Gini index.

SVM is one of the most widely used ML algorithm for dealing with binary classification problem and has been widely used in computational biology [13,27,29–31]. This method is based on Vapnik-

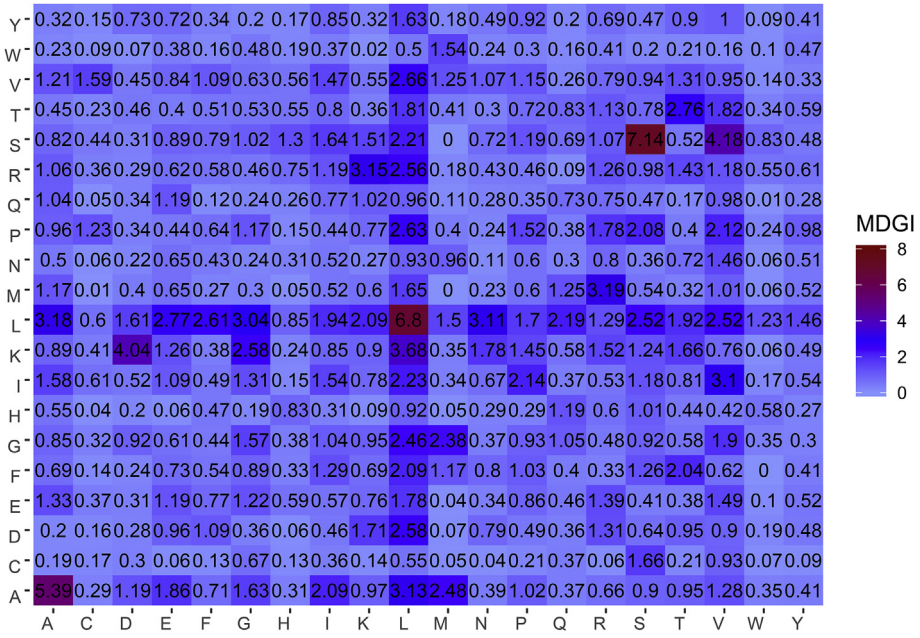

Fig. 2. Heat map of the mean decrease of Gini index (MDGI) of dipeptide compositions. It should be noted that features with the largest value of MDGI are deemed to be the most important.

Chervonenkis's theory of statistical learning [55–57]. This method determines the optimal hyperplane with the largest distance between two classes that minimizes the misclassification rate. To make linear separation on high dimensional samples, SVM employs one of the available well-known kernel functions to transform the sample space having *p*-dimensional feature vector into the feature space with *n*-dimensional feature vector, where $p < n$. In this work, the widely used radial basis function is applied to non-linearly transform the feature space. Herein, the RF and SVM classifiers were applied in the development of predictive models using the scikit-learn package (version 0.22) with default parameters.

2.5. Comparison between iTTCA-Hybrid with TTAgP1.0

To the best of our knowledge, there is only one study focusing on the prediction of tumor T cell antigens represented in the MHC class I context namely the TTAgP1.0 [39]. TTAgP1.0 is constructed using RF and two types of peptide features (i.e., Rfre and AAC). Since, the present study makes use of up-to-date dataset containing 592 tumor and 394 non-tumor T cell antigens, it is therefore not fair to directly compare the results of TTAgP1.0 with that of the proposed iTTCA-Hybrid. In order to make a fair comparison, a modified TTAgP1.0 method (TTAgP1.0-MODI) was established using the RF model with the two aforementioned feature types on our new dataset.

2.6. Performance evaluation

In order to evaluate the prediction ability of the model, we used four widely used metrics for solving the two-class prediction problem as follows:

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP}+\mathrm{TN} + \mathrm{FP}+\mathrm{FN})} \tag{5}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{6}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{7}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{8}$$

where Ac, Sn, Sp and MCC represents the accuracy, sensitivity, specificity and Matthews coefficient correlation, respectively. Meanwhile, TP is the number of true positives and TN is the number of true negatives, FP is the number of false positives and FP is the number of false negatives. To investigate the prediction performance of the proposed model using threshold-independent parameters, receiver operating characteristic (ROC) curves were plotted. The area under the ROC curve (auAUC) was utilized to assess the prediction performance where AUC values of 0.5 and 1 are indicative of random and perfect models, respectively [13,41,58–62].

3. Results and discussion

3.1. Analysis of amino acids and dipeptides preference

RF model provides two measures for ranking feature importance namely the mean decrease of the Gini index (MDGI) and the mean decrease of prediction accuracy. Since Calle and Urrea [63] demonstrated that the MDGI provided more robust results when compared with that of the mean decrease of prediction accuracy, we therefore use the MDGI to rank the importance of interpretable features that includes AAC and DPC. Features with the largest MDGI value was considered to be an important feature as it significantly contributed to the prediction performance. Supplementary Table S1 lists the percentage values of 20 amino acids for both tumor and non-tumor T cell antigens from the TTCA-CV dataset along with the amino acid composition difference between the two classes as well as their MDGI values. In addition, a heatmap showing the feature importance for DPC is provided in Fig. 2. Features with the largest MDGI value was considered to be an important feature [54]. As seen in Table 1, it can be seen that Glu and Pro are presented in high proportion in the tumor T cell antigens while Ser and Val are present in high proportions in the non-tumor T cell antigens. On the other hand, the top-five informative dipeptides consisted of SS, LL, AA, SV and KD with corresponding MDGI values of 7.138, 6.801, 5.390, 4.181 and 4.044, respectively, as shown in Fig. 2.

Table 2
Comparison of SVM and RF models with various types of features without an oversampling approach (SMOTE) over 10-fold cross-validation.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	70.56	88.30	44.35	0.369	0.721
	SVM	63.70	87.66	28.31	0.209	0.618
DPC	RF	66.11	81.28	43.74	0.274	0.686
	SVM	69.55	85.96	45.35	0.348	0.725
PseAAC	RF	66.76	82.55	43.43	0.285	0.689
	SVM	69.28	93.62	33.31	0.344	0.711
CTDD	RF	72.46	87.66	49.99	0.417	0.753
	SVM	70.05	92.55	36.77	0.365	0.689
PCP	RF	62.94	73.62	47.23	0.217	0.631
	SVM	60.78	91.49	15.42	0.120	0.567
DPC + PseAAC	RF	71.06	88.30	45.56	0.386	0.723
	SVM	68.91	87.23	41.85	0.331	0.725
DPC + CTDD	RF	71.82	88.51	47.14	0.400	0.756
	SVM	70.18	96.38	31.45	0.384	0.679
PseAAC + CTDD	RF	73.35	88.51	50.93	0.436	0.752
	SVM	71.32	93.19	38.98	0.399	0.688
DPC + PseAAC + CTDD	RF	72.84	89.36	48.41	0.425	0.759
	SVM	70.30	96.38	31.76	0.386	0.680

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

3.2. Evaluation of various feature extraction methods without SMOTE

In this section, we carried out comparative experiments via the use of five basic features (i.e., AAC, DPC, PseAAC, CTDD and PCP) and their hybrid features (i.e., DPC + PseAAC, DPC + CTDD, PseAAC + CTDD and DPC + PseAAC + CTDD) to assess their contributions to the prediction of tumor T-cell antigens without the use of the SMOTE algorithm. We hypothesized that using the combinations of various feature types could alleviate the weakness of one another. Each feature was assessed by using two robust ML classifiers namely SVM and RF and evaluated via 10-fold CV and independent test. Although, the jackknife test can provide unique results for any given training dataset but this comes at the cost of its time-consuming nature. Thus, the 10-fold CV and independent tests were used for evaluating the prediction performance of proposed models. Tables 2 and 3 list the results of performance comparisons of the comparative experiments over 10-fold CV and independent test, respectively. Furthermore, Fig. 3A and B plot ROC curves of the three selected optimal models over 10-fold CV and independent test, respectively. Several observations can be made from Tables 2 and 3 as follows.

First, in case of using the basic feature, the RF model making use of the CTDD feature achieved the highest Ac of 72.46% with MCC of 0.417 and AUC of 0.753, while the second highest Ac of 70.56% with MCC of 0.369 and AUC of 0.721 was obtained by RF model built using the AAC feature. The prediction performance of the RF model built with CTDD attained on an independent dataset was 70.56% Ac, 0.355 MCC and 0.739 AUC. It was observed that the CTDD feature could be effectively used for prediction of tumor T-cell antigens with Ac in excess of 70% as assessed by 10-fold CV. Secondly, in the case of using hybrid features, RF models built using hybrid features of PseAAC + CTDD and DPC + PseAAC + CTDD performed well with the first and second highest Ac of 73.35% and 72.74%, respectively. Interestingly, these two hybrid features consistently provided acceptable results with Ac of greater than 70% as evaluated on the independent dataset.

Table 3
Comparison of SVM and RF models with various types of features without an oversampling approach (SMOTE) over independent test.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	69.54	81.97	49.33	0.332	0.724
	SVM	65.48	84.43	34.67	0.221	0.693
DPC	RF	59.90	73.77	37.33	0.117	0.658
	SVM	69.54	84.43	45.33	0.326	0.780
PseAAC	RF	67.51	80.33	46.67	0.286	0.721
	SVM	72.59	95.08	36.00	0.404	0.756
CTDD	RF	70.56	82.79	50.67	0.355	0.739
	SVM	75.63	93.44	46.67	0.471	0.786
PCP	RF	63.96	72.95	49.33	0.226	0.637
	SVM	58.88	86.07	14.67	0.010	0.524
DPC + PseAAC	RF	68.02	83.61	42.67	0.289	0.713
	SVM	72.59	86.89	49.33	0.397	0.782
DPC + CTDD	RF	70.05	83.61	48.00	0.340	0.747
	SVM	76.65	97.54	42.67	0.511	0.792
PseAAC + CTDD	RF	71.07	84.43	49.33	0.363	0.756
	SVM	75.13	93.44	45.33	0.460	0.785
DPC + PseAAC + CTDD	RF	72.08	86.07	49.33	0.385	0.757
	SVM	76.65	97.54	42.67	0.511	0.792

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

3.3. Evaluation of various features extraction methods with SMOTE

In this section, we attempt to demonstrate the utility of the oversampling approach (SMOTE) for enhancing the prediction performance by treating the data imbalance problem. In order to evaluate the impact of the oversampling approach, the same comparative experiments using five basic features and their hybrid features were carried out. Tables 3 and 4 lists the results from performance comparisons as evaluated via 10-fold CV and independent test, respectively. Furthermore, Fig. 3C and D shows the plot of ROC curves for the three selected optimal models as assessed by 10-fold CV and independent test, respectively. Several observations can be made as follows.

By comparing Tables 4 and 5, it could be clearly noticed that prediction results of classifiers were improved by using the oversampling approach. Firstly, in case of using the basic feature, the highest Ac values of 77.55%, 76.70 and 76.17% over 10-fold CV were achieved by using RF with AAC, RF with CTDD and SVM with DPC features, respectively. Secondly, in case of utilizing the hybrid features, RF models built in conjunction with hybrid features of DPC + CTDD and PseAAC + CTDD provided Ac of 78.83% and 77.45%, respectively. Remarkably, using the hybrid feature of PseAAC + CTDD can efficaciously predict tumor T-cell antigens with an Ac of 73.35%, MCC of 0.436 and AUC of 0.752 on the independent dataset. Based on the observation that can be deduced from Tables 2–5, the optimal prediction performance as evaluated by 10-fold CV and independent dataset were achieved by using RF as the predictor together with the use of the hybrid feature of PseAAC + CTDD as the input feature on a balanced dataset derived from the SMOTE method. For convenience of the subsequent description, we will refer to this method as iTTCA-Hybrid.

3.4. Reliability of iTTCA-Hybrid

To avoid the bias of single random sampling, the dataset was subjected to 20 iterations of randomly generated training and independent datasets [12,17,18,49,50]. Thus, 20 prediction models were built and assessed via 10-fold CV and independent tests. Prediction results are provided in Supplementary Table S2. It can be seen from Supplementary Table S2 that iTTCA-Hybrid yields accuracy and MCC of 78.73 ± 1.18 and 0.851 ± 0.012, respectively, as assessed by 10-fold CV. Meanwhile, independent test results of the predictor presented herein afforded accuracy and MCC of 73.65 ± 1.14 and 0.437 ± 0.032, respectively. Prediction results indicated that the proposed iTTCA-Hybrid is accurate, reliable and stable for the identification of tumor T cell antigens.

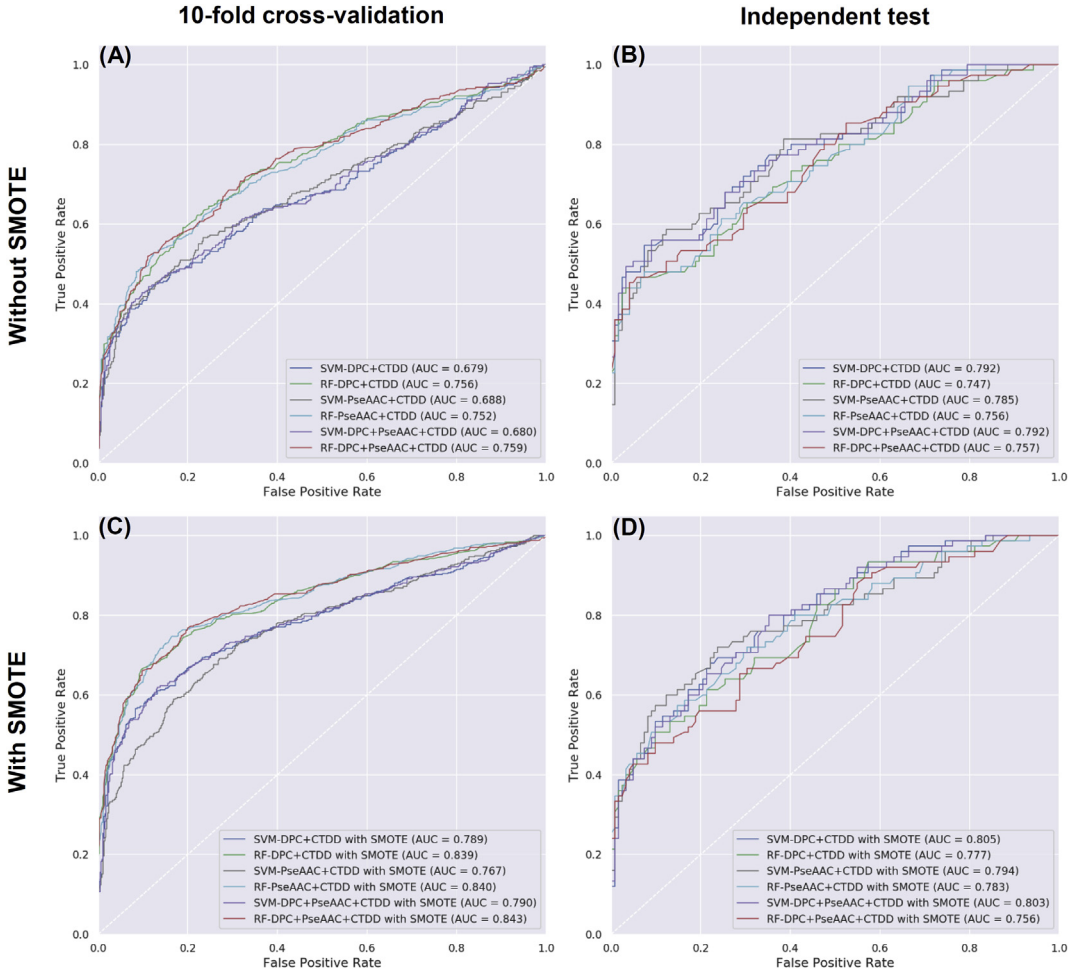

Fig. 3. The ROC curves of RF and SVM models in conjunction with various hybrid features over 10-fold cross-validation (A and C) and independent test (B and D), and with (C and D) and without (A and B) SMOTE method.

Table 4
Comparison of SVM and RF models with various types of features with an oversampling approach (SMOTE) over 10-fold cross-validation.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	77.55	86.60	68.51	0.573	0.839
	SVM	70.64	71.91	69.36	0.423	0.765
DPC	RF	73.30	79.36	67.23	0.490	0.830
	SVM	76.17	82.98	69.36	0.544	0.851
PseAAC	RF	75.53	81.28	69.79	0.526	0.824
	SVM	69.89	84.89	54.89	0.417	0.750
CTDD	RF	76.70	84.89	68.51	0.546	0.834
	SVM	70.53	78.94	62.13	0.418	0.768
PCP	RF	67.77	70.00	65.53	0.362	0.733
	SVM	54.89	67.66	42.13	0.103	0.583
DPC + PseAAC	RF	75.85	85.11	66.60	0.537	0.835
	SVM	76.38	84.26	68.51	0.550	0.849
DPC + CTDD	RF	77.45	86.81	68.09	0.566	0.839
	SVM	72.87	83.40	62.34	0.471	0.789
PseAAC + CTDD	RF	78.83	85.53	72.13	0.588	0.840
	SVM	70.43	79.36	61.49	0.417	0.767
DPC + PseAAC + CTDD	RF	77.13	85.11	69.15	0.555	0.843
	SVM	73.62	84.04	63.19	0.486	0.790

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

Table 5
Comparison of SVM and RF models with various types of features with an oversampling approach (SMOTE) over independent test.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	67.01	78.69	48.00	0.279	0.716
	SVM	68.53	72.13	62.67	0.343	0.706
DPC	RF	65.48	77.87	45.33	0.244	0.702
	SVM	69.54	81.97	49.33	0.332	0.775
PseAAC	RF	68.02	77.87	52.00	0.307	0.736
	SVM	67.51	78.69	49.33	0.292	0.761
CTDD	RF	68.53	77.05	54.67	0.323	0.746
	SVM	75.63	82.79	64.00	0.476	0.800
PCP	RF	63.96	71.31	52.00	0.234	0.655
	SVM	56.35	66.39	40.00	0.065	0.541
DPC + PseAAC	RF	69.04	80.33	50.67	0.324	0.737
	SVM	69.04	80.33	50.67	0.324	0.775
DPC + CTDD	RF	72.08	82.79	54.67	0.392	0.777
	SVM	74.11	86.07	54.67	0.434	0.805
PseAAC + CTDD	RF	73.60	82.79	58.67	0.428	0.783
	SVM	75.13	83.61	61.33	0.462	0.794
DPC + PseAAC + CTDD	RF	72.08	81.97	56.00	0.394	0.756
	SVM	73.60	84.43	56.00	0.425	0.803

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

Table 6
Comparison of iTTCA-Hybrid with other related methods over 10-fold cross-validation and independent test.

Cross-validation	Classifier[a]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
10-fold CV	TTAgP1.0-MODI[a]	70.81	86.17	48.09	0.375	0.730
	iTTCA-Hybrid[a]	73.35	88.51	50.93	0.436	0.752
	TTAgP1.0-MODI[b]	77.34	83.83	70.85	0.565	0.838
	iTTCA-Hybrid[b]	78.83	85.53	72.13	0.588	0.840
Independent test	TTAgP1.0-MODI[a]	69.54	81.97	49.33	0.332	0.737
	iTTCA-Hybrid[a]	71.07	84.43	49.33	0.363	0.756
	TTAgP1.0-MODI[b]	71.07	78.69	58.67	0.379	0.747
	iTTCA-Hybrid[b]	73.60	82.79	58.67	0.428	0.783

[a] Models were performed on unbalanced data.
[b] Models were performed on balanced data using an oversampling approach (SMOTE).

3.5. Comparison of iTTCA-Hybrid with TTAgP1.0

To further assess the predictive efficiency and effectiveness of the proposed iTTCA-Hybrid, the model was subjected to comparison with existing method (TTAgP1.0). Herein, we developed the TTAgP1.0-MODI to fairly compare the proposed model by performing the same experimental setting and cross-validation methods. Table 6 and Fig. 4 show the performance comparison between iTTCA-Hybrid and TTAgP1.0-MODI as evaluated by 10-fold CV and independent test.

In the case of 10-fold CV test, the Ac, MCC and AUC of iTTCA-Hybrid as performed on the imbalanced dataset afforded performance values of 73.35%, 0.436 and 0.752, respectively, while the TTAgP1.0-MODI model yielded 70.81%, 0.375 and 0.732, respectively. As can be noticed in Table 6 and Fig. 4, the iTTCA-Hybrid model that make use of the SMOTE algorithm for oversampling was found to outperform TTAgP1.0-MODI with improvements of 8%, 10% and 9%, respectively, for Ac, MCC and AUC. As for the independent test (Table 6), similar results to those obtained from the cross-validation test was observed. Particularly, iTTCA-Hybrid was found to achieve significantly improved performance than that of TTAgP1.0-MODI as assessed by the five aforementioned statistical parameters. Interestingly, iTTCA-Hybrid was found to afford higher performance than that of TTAgP1.0-MODI in which Ac, Sp, MCC and AUC demonstrated improvements of 4%, 9%, 9% and 6%, respectively. Taken all together, comparative results demonstrated that the predictor proposed herein was able to achieve significantly improved as compared to those of existing method, TTAgP1.0-MODI.

3.6. iTTCA-hybrid web server

In order to facilitate rapid and easy use of the predictive model presented herein as well as for the benefit of the scientific community, we herein developed a user-friendly web server known as the iTTCA-Hybrid, which is made freely available online at http://camt.pythonanywhere.com/iTTCA-Hybrid. Step-by-step guidelines on how to use the iTTCA-Hybrid web server in order to obtain prediction results are as follows. Firstly, the user accesses the web server at http://camt.pythonanywhere.com/iTTCA-Hybrid where the user will see the top screen of iTTCA-Hybrid as shown in Fig. 5. Secondly, the user enters the query sequence into the text box or uploads a FASTA file by clicking on the “Choose file” button. Thirdly, the user clicks on the “Submit” button in order to start the prediction process. Typically, in this step, it takes a few seconds for the server to process the task. Finally, after finishing the prediction process, results are displayed as shown on the right-hand side of the web server. The user can see examples of FASTA-formatted sequences by clicking on the “example file” button.

4. Discussion

The aim of the present study is to develop an accurate, unbiased and automated sequence-based tumor T cell antigen predictor named herein as the iTTCA-Hybrid. To the best of our knowledge, there is only one study in existence that focuses on the prediction of tumor T cell antigens represented in the MHC class I context namely the TTAgP1.0 [39]. As mentioned above, our comparative analyses revealed that the method proposed herein have been found to afford improved performance and reliability. Herein, we further discuss about the prediction as summarized in the following paragraphs.

The first point is the objectivity of the benchmark dataset. So as to afford a fair comparison, the same benchmark dataset should be used to develop and assess the proposed model in order to provide a comprehensive and unbiased comparison [64–68]. However, previous studies did not publicly share the dataset used for the model development [39]. To address this issue, we compiled a benchmark dataset containing 529 experimentally verified tumor T cell antigens from the TANTIGEN [42] and TANTIGEN 2.0 [43] databases as well as 393 non-tumor T cell antigens collected from IEDB database (https://www.iedb.org/).

The second point pertains to the significance of using various feature types for covering various aspects of sequence information [20,22,23,28]. TTAgP1.0 was developed by making use of only two types of peptide features namely the Rfre and PEP. In the meantime, the proposed iTTCA-Hybrid was built using five basic features and their hybrid features. Table 3 shows that the highest accuracy of 73.35% was achieved from the hybrid feature of PseAAC + CTDD, while TTAgP1.0

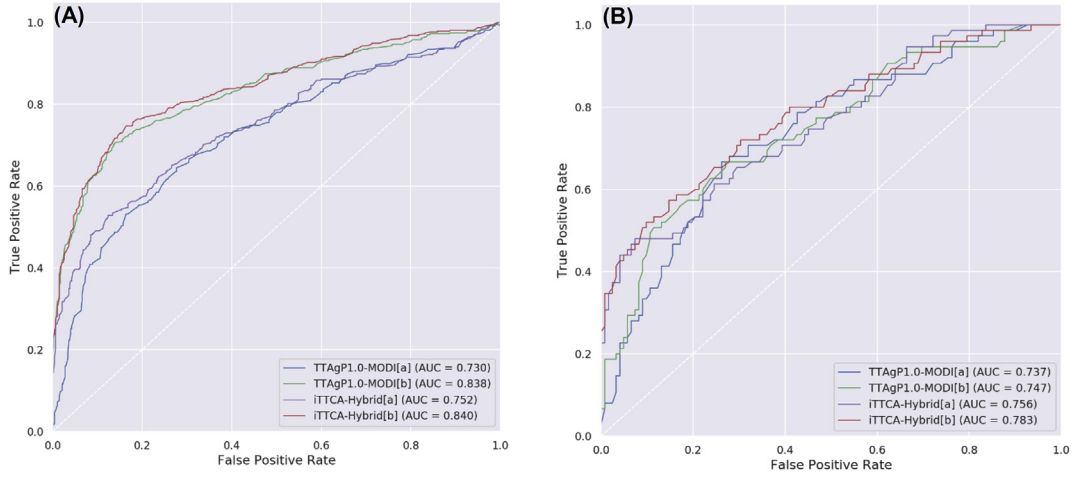

Fig. 4. The ROC curves of iTTCA-Hybrid with TTAgP1.0 over 10-fold cross-validation (A) and independent test (B).

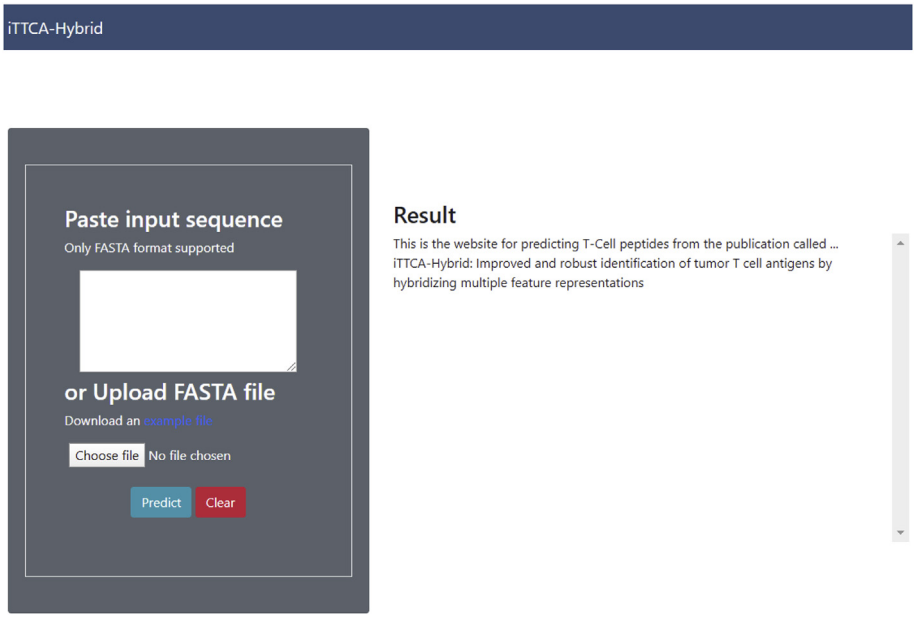

Fig. 5. Screenshots of the iTTCA-Hybrid web server.

afforded lower accuracy of 70.81%. Results indicated that the hybrid feature described herein was more beneficial for identifying tumor T cell antigen than that of Rfre and PEP.

The third point concerns the problem of bias arising from the class imbalance problem. Since, our benchmark dataset was imbalanced where the ratio of non-tumor to tumor T cell antigens was 1 : 1.35. From the machine learning point of view, the class imbalance dataset has the tendency to cause erroneous prediction model that may potentially be overfitting as well as perform poorly on the minority class [6,45]. To overcome this mentioned problem, the SMOTE method was utilized to cope with the potential bias arising from the class imbalance problem [6,45,50–52]. Based on our comparative results, we observed that iTTCA-Hybrid was superior to that of TTAgP1 as evaluated by Ac, Sp, MCC and AUC in which improvements of 4%, 9%, 9% and 6%, respectively, were observed when compared to that of TTAgP1.0. These results indicated that our proposed model that makes use of the SMOTE method were more efficient and reliable in this aspect.

The fourth point deals with the importance of the deploying the predictive model as a publicly web server. The benefit of hosting the model as a web server are many namely facilitating the access of the predictive capabilities of the model by experimental biologists without the need to develop in-house models [19,21,30,69–72]. No web server was provided by previous studies, which is therefore addressed in the present study whereby the iTTCA-Hybrid web server was established and made freely available online at http://camt.pythonanywhere.com/iTTCA-Hybrid. This implemented web server allows fast and easy robust predictions to be made without the need to develop in-house prediction models.

5. Conclusions

In light of the promise of cancer immunotherapy, we have developed a reliable, accurate, unbiased and automated sequence-based tumor T cell antigen predictor named iTTCA-Hybrid with high prediction performance. This is implemented by effectively employing multiple peptide features together with the use of the robustness of random forest. Benchmarking results affirmed that the proposed iTTCA-Hybrid produced significant performance improvements as compared to those of existing methods. Finally, for the convenience of experimental biologists, the iTTCA-Hybrid model was established as freely available web server that can be accessed at http://camt.pythonanywhere.com/iTTCA-Hybrid. It is highly anticipated that the proposed iTTCA-Hybrid represents a useful and robust tool for rational screening and design of cancer vaccines. Although, the predictor presented herein afforded the best performance than those of existing methods, there is still ample room for further improvements. For instance, ongoing work in our lab is geared towards improving the usefulness and applicability for field of cancer immunotherapy via the use of the interpretable scoring card method (SCM) [31,40,41] for shedding insights on the biophysical and biochemical properties of tumor T cell antigens.

Author contributions

W.S. conceived and analyzed the experiments. W.S. and P.C. designed and performed the experiments. P.C. and W.S analyzed the data. P.C contributed the code for constructing iTTCA-Hybrid model and the web server. W.S., C.N., M.H. and P.C. drafted the manuscript. All authors read and approved the manuscript.

Declaration of competing interest

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

Acknowledgments

This work was fully supported by the TRF Research Grant for New Scholar (No. MRG6180226) and College of Arts, Media and Technology, Chiang Mai University, and partially supported by the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund, the Office of Higher Education Commission and Mahidol University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ab.2020.113747.

References

[1] V.O. Carvalho, Left atrial volume and exercise capacity in adult heart transplant recipients, J. Cardiothorac. Surg. 6 (1) (2011) 9.

[2] S. Simard, et al., Fear of cancer recurrence in adult cancer survivors: a systematic review of quantitative studies, J. Canc. Survivorship 7 (3) (2013) 300–322.
[3] H. Perdry, B.S. Maher, M.-C. Babron, T. McHenry, F. Clerget-Darpoux, M.L. Marazita, An ordered subset approach to including covariates in the transmission disequilibrium test, BMC Proc. 1 (1) (2007) S77 BioMed Central.
[4] J. Couzin-Frankel, Cancer Immunotherapy, American Association for the Advancement of Science, 2013.
[5] H. Li, N. Anuwongcharoen, A.A. Malik, V. Prachayasittikul, J.E. Wikberg, C. Nantasenamat, Roles of d-amino acids on the bioactivity of host defense peptides, Int. J. Mol. Sci. 17 (7) (2016) 1023.
[6] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, ACPred: a computational tool for the prediction and analysis of anticancer peptides, Molecules 24 (10) (2019) 1973.
[7] H. Li, N. Schaduangrat, S. Simeon, C. Nantasenamat, Computational study on the origin of the cancer immunotherapeutic potential of B and T cell epitope peptides, Mol. Biosyst. 13 (11) (2017) 2310–2322.
[8] E. Mizukoshi, et al., Comparative analysis of various tumor-associated antigen-specific t-cell responses in patients with hepatocellular carcinoma, Hepatology 53 (4) (2011) 1206–1216.
[9] T.N. Schumacher, R.D. Schreiber, Neoantigens in cancer immunotherapy, Science 348 (6230) (2015) 69–74.
[10] S. Bobisse, P.G. Foukas, G. Coukos, A. Harari, Neoantigen-based cancer immunotherapy, Ann. Transl. Med. 4 (14) (2016).
[11] S.K. Saini, N. Rekers, S.R. Hadrup, Novel tools to assist neoepitope targeting in personalized cancer immunotherapy, Ann. Oncol. 28 (suppl_12) (2017) xii3–xii10.
[12] T.S. Win, A.A. Malik, V. Prachayasittikul, J.E. S Wikberg, C. Nantasenamat, W. Shoombuatong, HemoPred: a web server for predicting the hemolytic activity of peptides, Future Med. Chem. 9 (3) (2017) 275–291.
[13] M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, B. Manavalan, HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation, Bioinformatics (2020) btaa160.
[14] B. Manavalan, R.G. Govindaraj, T.H. Shin, M.O. Kim, G. Lee, iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction, Front. Immunol. 9 (2018) 1695.
[15] S. Basith, B. Manavalan, T.H. Shin, G. Lee, iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree, Comput. Struct. Biotechnol. J. 16 (2018) 412–420.
[16] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation, Int. J. Mol. Sci. 20 (22) (2019) 5743.
[17] T.S. Win, N. Schaduangrat, V. Prachayasittikul, C. Nantasenamat, W. Shoombuatong, PAAP: a web server for predicting antihypertensive activity of peptides, Future Med. Chem. 10 (15) (2018) 1749–1767.
[18] W. Shoombuatong, N. Schaduangrat, R. Pratiwi, C. Nantasenamat, THPep: a machine learning-based approach for predicting tumor homing peptides, Comput. Biol. Chem. 80 (2019) 441–451.
[19] B. Manavalan, S. Basith, T.H. Shin, S. Choi, M.O. Kim, G. Lee, MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 8 (44) (2017) 77121.
[20] B. Manavalan, S. Basith, T.H. Shin, D.Y. Lee, L. Wei, G. Lee, 4mCpred-EL: An ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome, Cells 8 (11) (2019) 1332.
[21] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees, Comput. Struct. Biotechnol. J. 17 (2019) 972–981.
[22] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation, Mol. Ther. Nucleic Acids 16 (2019) 733–744.
[23] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, Bioinformatics 35 (16) (2019) 2757–2765.
[24] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest, Front. Pharmacol. 9 (2018) 276.
[25] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions, Front. Immunol. 9 (2018) 1783.
[26] B. Manavalan, T.H. Shin, G. Lee, PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine, Front. Microbiol. 9 (2018) 476.
[27] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i4mC-Mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes, Comput. Struct. Biotechnol. J. 18 (2020) 906–912.
[28] M.M. Hasan, B. Manavalan, M.S. Khatun, H. Kurata, i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome, Int. J. Biol. Macromol. (2019), https://doi.org/10.1016/j.ijbiomac.2019.12.009.
[29] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation, Plant Mol. Biol. (2020) 1–10.
[30] W. Chen, P. Feng, F. Nie, iATP: A Sequence Based Method for Identifying Anti-tubercular Peptides, Medicinal chemistry (Shariqah (United Arab Emirates)), 2019.
[31] P. Charoenkwan, W. Shoombuatong, H.-C. Lee, J. Chaijaruwanich, H.-L. Huang, S.-Y. Ho, SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs, PloS One 8 (9) (2013).
[32] W. Shoombuatong, S. Hongjaisee, F. Barin, J. Chaijaruwanich, T. Samleerat, HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees, Comput. Biol. Med. 42 (9) (2012) 885–889.
[33] S. Hongjaisee, C. Nantasenamat, T.S. Carraway, W. Shoombuatong, HIVCoR: a sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage, Comput. Biol. Chem. 80 (2019) 419–432.
[34] X. Yang, X. Yu, An introduction to epitope prediction methods and software, Rev. Med. Virol. 19 (2) (2009) 77–96.
[35] M. Bhasin, G. Raghava, Prediction of CTL epitopes using QM, SVM and ANN techniques, Vaccine 22 (23–24) (2004) 3195–3204.
[36] K. Yu, N. Petrovsky, C. Schönbach, J.L. Koh, V. Brusic, Methods for prediction of peptide binding to MHC molecules: a comparative study, Mol. Med. 8 (3) (2002) 137–148.
[37] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, L. Harrison, Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, Bioinformatics 14 (2) (1998) 121–130.
[38] J.C. Tong, T.W. Tan, S. Ranganathan, Modeling the structure of bound peptide ligands to major histocompatibility complex, Protein Sci. 13 (9) (2004) 2523–2532.
[39] J.F.B. Lissabet, L.H. Belén, J.G. Farias, TTAgP 1.0: a computational tool for the specific prediction of tumor T cell antigens, Comput. Biol. Chem. 83 (2019) 107103.
[40] P. Charoenkwan, S. Kanthawong, N. Schaduangrat, J. Yana, W. Shoombuatong, PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method, Cells 9 (2) (2020) 353.
[41] P. Charoenkwan, J. Yana, N. Schaduangrat, C. Nantasenamat, M.M. Hasan, W. Shoombuatong, iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides, Genomics (2020), https://doi.org/10.1016/j.ygeno.2020.03.019.
[42] L.R. Olsen, S. Tongchusak, H. Lin, E.L. Reinherz, V. Brusic, G.L. Zhang, TANTIGEN: a comprehensive database of tumor T cell antigens, Canc. Immunol. Immunother. 66 (6) (2017) 731–735.
[43] G. Zhang, L. Chitkushev, D.B. Keskin, V. Brusic, TANTIGEN 2.0: an online database and analysis platform for tumor T cell antigens, 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 2228–2231.
[44] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (1) (2000) 374-374.
[45] R. Pratiwi, et al., CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins, J. Chem. 2017 (2017).
[46] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (3) (2015) 218–234.
[47] K.-C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247.
[48] Z. Chen, et al., iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (14) (2018) 2499–2502.
[49] V. Laengsri, C. Nantasenamat, N. Schaduangrat, P. Nuchnoi, V. Prachayasittikul, W. Shoombuatong, TargetAntiAngio: a sequence-based tool for the prediction and analysis of anti-angiogenic peptides, Int. J. Mol. Sci. 20 (12) (2019) 2950.
[50] X. He, et al., TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition, J. Membr. Biol. 248 (6) (2015) 1005–1014.
[51] E. Ramentol, Y. Caballero, R. Bello, F. Herrera, SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, Knowl. Inf. Syst. 33 (2) (2012) 245–265.
[52] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[53] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[54] L. Breiman, Classification and Regression Trees, Routledge, 2017.
[55] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 2013.
[56] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[57] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Network. 10 (5) (1999) 988–999.
[58] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, L. Hao, Predicting protein structural classes for low-similarity sequences by evaluating different features, Knowl. Base Syst. 163 (2019) 787–793.
[59] H. Lin, Z.-Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE ACM Trans. Comput. Biol. Bioinf 16 (4) (2017) 1316–1321.
[60] H.-Y. Lai, et al., iProEP: a computational predictor for predicting promoter, Mol. Ther. Nucleic Acids 17 (2019) 337–346.
[61] C.-Q. Feng, et al., iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators, Bioinformatics 35 (9) (2019) 1469–1477.
[62] F.-Y. Dao, et al., Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique, Bioinformatics 35 (12) (2019) 2075–2083.
[63] M.L. Calle, V. Urrea, Letter to the editor: stability of random forest importance measures, Briefings Bioinf. 12 (1) (2010) 86–89.
[64] W. Shoombuatong, et al., Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling, Curr. Drug Metabol. 18 (6) (2017) 540–555.
[65] N. Schaduangrat, S. Lampa, S. Simeon, M.P. Gleeson, O. Spjuth, C. Nantasenamat, Towards reproducible computational drug discovery, J. Cheminf. 12 (1) (2020) 9.
[66] W. Shoombuatong, et al., Towards the revival of interpretable QSAR models, Advances in QSAR Modeling, Springer, 2017, pp. 3–55.
[67] W. Shoombuatong, N. Schaduangrat, C. Nantasenamat, Towards understanding aromatase inhibitory activity via QSAR modeling, EXCLI J. 17 (2018) 688.
[68] W. Shoombuatong, N. Schaduangrat, C. Nantasenamat, Unraveling the bioactivity of anticancer peptides as deduced from machine learning, EXCLI J. 17 (2018) 734.

[69] M.M. Hasan, M.M. Rashid, M.S. Khatun, H. Kurata, Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information, Sci. Rep. 9 (1) (2019) 1–9.
[70] M.M. Hasan, S. Yang, Y. Zhou, M.N.H. Mollah, SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties, Mol. Biosyst. 12 (3) (2016) 786–795.
[71] M.M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song, Z. Zhang, Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs, PloS One 10 (6) (2015).
[72] K. Liu, W. Chen, iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications, Bioinformatics (2020), https://doi.org/10.1093/bioinformatics/btaa155.