RESEARCH

Open Access

iTTCA-RF: a random forest predictor for tumor T cell antigens

Shihu Jiao[1], Quan Zou[1,2], Huannan Guo[3*] and Lei Shi[4*]

Abstract

Background: Cancer is one of the most serious diseases threatening human health. Cancer immunotherapy represents the most promising treatment strategy due to its high efficacy and selectivity and lower side effects compared with traditional treatment. The identification of tumor T cell antigens is one of the most important tasks for antitumor vaccines development and molecular function investigation. Although several machine learning predictors have been developed to identify tumor T cell antigen, more accurate tumor T cell antigen identification by existing methodology is still challenging.

Methods: In this study, we used a non-redundant dataset of 592 tumor T cell antigens (positive samples) and 393 tumor T cell antigens (negative samples). Four types feature encoding methods have been studied to build an efficient predictor, including amino acid composition, global protein sequence descriptors and grouped amino acid and peptide composition. To improve the feature representation ability of the hybrid features, we further employed a two-step feature selection technique to search for the optimal feature subset. The final prediction model was constructed using random forest algorithm.

Results: Finally, the top 263 informative features were selected to train the random forest classifier for detecting tumor T cell antigen peptides. iTTCA-RF provides satisfactory performance, with balanced accuracy, specificity and sensitivity values of 83.71%, 78.73% and 88.69% over tenfold cross-validation as well as 73.14%, 62.67% and 83.61% over independent tests, respectively. The online prediction server was freely accessible at http://lab.malab.cn/~acy/iTTCA.

Conclusions: We have proven that the proposed predictor iTTCA-RF is superior to the other latest models, and will hopefully become an effective and useful tool for identifying tumor T cell antigens presented in the context of major histocompatibility complex class I.

Keywords: Tumor T cell antigens, Random forest, MRMD, Feature selection, Hybrid features

Introduction

According to a report from the International Agency for Research on Cancer (IARC), approximately 10 million people die of cancer, and there were 19.3 million new cancer cases worldwide in 2020. Cancer has become the second leading cause of death [1, 2]. Tumor molecular targeted therapy, radiotherapy and chemotherapy together constitute the main means of modern cancer drug therapy [3–6]. Classic broad-spectrum anticancer drugs and radiotherapy are lethal to tumor cells, but they can destroy normal cells in the body, produce large adverse reactions, and are prone to drug resistance [7–11]. Advances in equipment and immuno-oncology are driving a revolution in the field of cancer care. New cancer treatments are emerging, and targeted

---
*Correspondence: guohuannan1226@sohu.com; slspine@163.com
[3] Department of Oncology, General Hospital of Heilongjiang Province Land Reclamation Bureau, Harbin, China
[4] Department of Spine Surgery, Changzheng Hospital, Naval Medical University, Shanghai, China
Full list of author information is available at the end of the article

Jiao *et al. J Transl Med*   (2021) 19:449

Page 2 of 11

immunotherapy is one of the most promising treatment options. Unlike the harmful side effects of chemotherapy and radiotherapy, immunotherapy has been proven to be highly selective and effective, while also reducing side effects [12–15]. Immunotherapy provides new opportunities for the development of potential cancer treatments. T cells can recognize and kill tumor antigens encountered on the surface, which are presented by major histocompatibility complex (MHC) class I and class II molecules on the antigen-presenting cell surface [16–18]. Therefore, T cells play an important role in the field of tumor rejection and immunotherapeutic cancer. Correctly identifying T cell antigens not only helps to understand their protective mechanism but also contributes to the development of highly efficient cancer peptide vaccines [19].

Although the experimental methods are considered to be the most reliable method to characterize the biological activity of T cell epitopes in tumor antigens, they are usually time-consuming and expensive. Due to their convenience and high efficiency, computational methods have attracted increasing attention in the field of bioinformatics [20–30]. In this study, we focused on the identification of tumor T cell antigens (TTCAs) represented by MHC class I. According to our research, only two machine learning prediction tools have been published to identify this type of TTCA. The first prediction model was introduced by Lissabet et al. and is called TTAgP1.0 [31]. TTAgP1.0 uses the relative frequency of amino acids and amino acid composition (AAC) to encode the peptide sequences and then employs random forest (RF) classifier to build the prediction model [32]. Regrettably, TTAgP1.0 neither provides a web-server nor the dataset used. Therefore, its usage for the related research community is quite limited, although it has its own advantages and reasonable prediction accuracy. Very recently, Charoenkwan et al. proposed another random forest based prediction model iTTCA-Hybrid [33]. Five feature extraction methods, namely, AAC, pseudo amino acid composition (PAAC), dipeptide composition (DPC), amino acid property distribution (CTDD) and physicochemical (PCP), were investigated. The final model was constructed using the hybrid features of PAAC and CTDD. In addition, the oversampling technique was also applied to address the problem of data imbalance.

In this paper, we present a new predictor, iTTCA-RF, to distinguish TTCA from non-TTCA more accurately. As shown in Fig. 1, the protein sequences were preliminarily encoded using four kinds of feature extraction methods, namely global protein sequence descriptors (GPSD), grouped amino acid and peptide composition (GAAPC), PAAC and adaptive skip dipeptide composition (ASDC). We have investigated the performance of

four single descriptors and their all possible combinations on six commonly classifiers, where the imbalanced training samples were handled by the hybrid-sampling approach SMOTE-Tomek. The results suggest that the hybrid feature composed of GPSD, GAAC and PAAC was the most informative for TTCA identification. Then, the maximum relevance maximum distance (MRMD) algorithm was used to analyze the feature importance of the involved vectors. With the application of the incremental feature selection (IFS) strategy, different feature subsets are generated for optimization under consideration of the classification algorithms. Ultimately, the best performance model was finally constructed using the top 263 selected features. The tenfold cross-validation (CV) scores of iTTCA-RF were balanced accuracy (BACC) = 83.71%, MCC = 0.678, AUC = 0.894, Sn = 88.69% and Sp = 78.73%, while those of the latest iTTCA-Hybrid were BACC = 78.83%, MCC = 0.588, AUC = 0.840, Sn = 85.53%, Sp = 72.13%. The iTTCA-RF achieved scores with BACC = 73.14%, MCC = 0.474 and Sp = 62.67% over the independent test, which means relative improvements of 2.4%, 4.6% and 4.0%, respectively, compared to the existing state-of-the-art model. We also established a user-friendly web server, which is expected to be an effective and useful tool for TTCA identification.

Materials and methods
Datasets
In this research, we directly used the benchmark datasets collected by Charoenkwan et al. [33]. The dataset was constructed as follows: (1) a total of 727 MHC class I peptides were collected as positive samples from TANTIGEN [34] and TANTIGEN 2.0 [12]; (2) non-TTCA was collected from the IEDB database [35], in addition, samples with no relationship with any disease were chosen as negative samples; and (3) duplicate peptide sequences were eliminated. Ultimately, 592 positive and 393 negative samples were obtained. As shown in Table 1, 80% of the samples were randomly selected as the training dataset and while the remaining 20% of samples as the independent test datasets.

Feature representation
The quality of extracted sample features will greatly affect the performance of the predictive model. Researchers have proposed various biological sequence encoding strategies that can conveniently convert protein sequences into numerical vectors [22, 36–46]. In this paper, four feature encoding methods described below were adopted to represent the peptide sequences. In this work, we used the iLearn tool package [37] to generate the four type of sequence features.
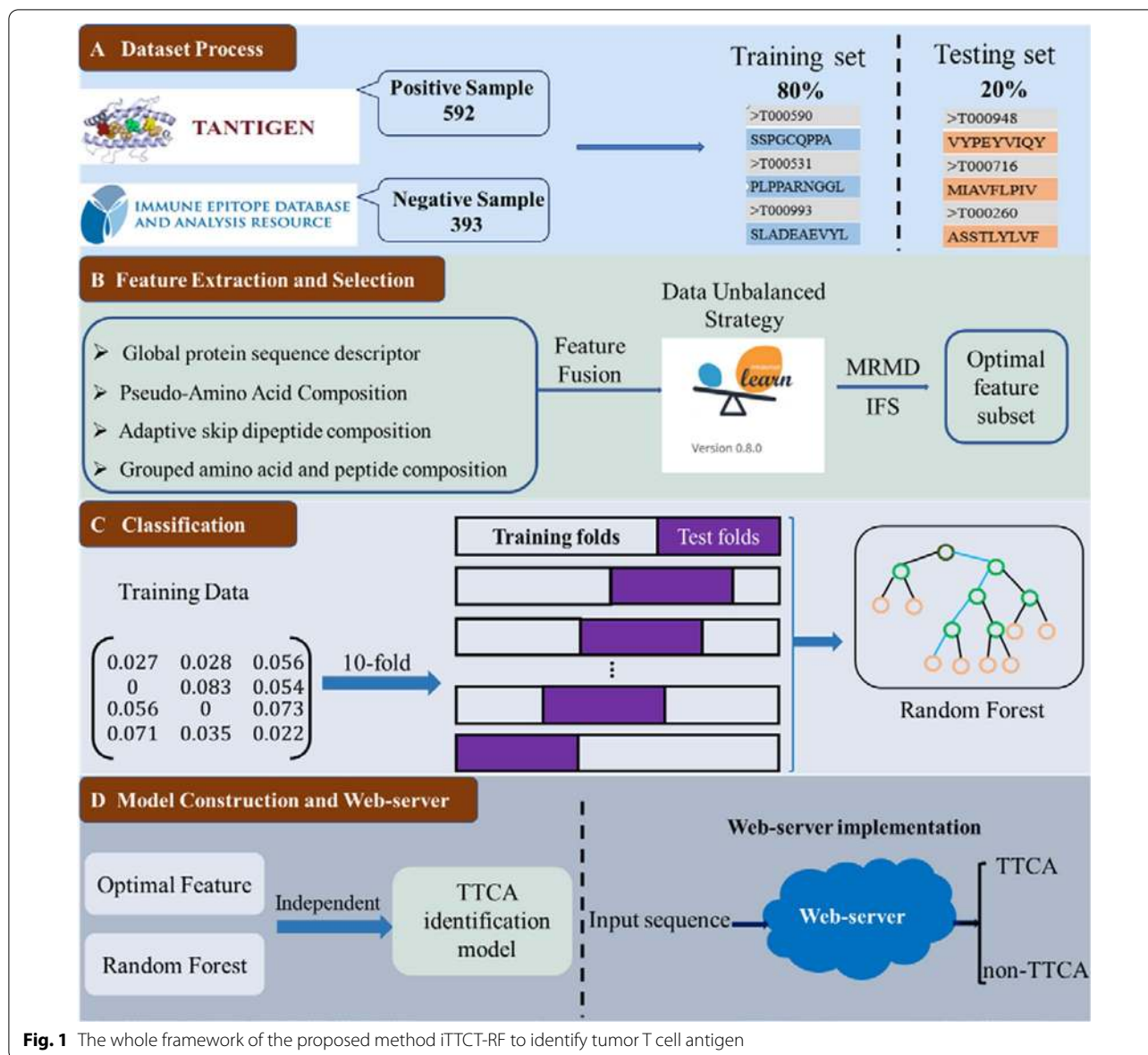
Jiao et al. J Transl Med   (2021) 19:449

Page 3 of 11

Fig. 1 The whole framework of the proposed method iTTCT-RF to identify tumor T cell antigen

Table 1 Sample distribution in the training and independent test datasets

Dataset	Training	Testing
Positive	470	122
Negative	318	75

Global protein sequence descriptor (GPSD)

This method (also called 188D features in many studies) describes the global composition of amino acid properties in a protein sequence and generates 188 features that integrate both sequence information and amino acid properties [47, 48]. In general, the GPSD descriptor contains

two parts. The first part is the amino acid composition. The amino acid frequency in the peptide was calculated to obtain the first 20 features. The second part is the 168 features related to eight physicochemical properties of amino acids. Detailed information about the eight physicochemical properties of amino acids is described in the References [49–51]. For each property, 20 amino acids were divided into three groups, and the CTD (C: composition, T: transition and D: distribution) pattern was applied to encode the peptide sequences to generate 21D features. C stands for the occurrence frequencies of each group (3D). T represents the transition frequencies between the three groups (3D). D refers to the first, 25%, 50%, 75% and last occurring positions of a certain group in the peptide

Jiao *et al. J Transl Med*   (2021) 19:449

Page 4 of 11

sequence ($5 \times 3 = 15D$). Thus, the CTD model will produce $8 * (3 + 3 + 15) = 168$ features.

Grouped amino acid and peptide composition (GAAPC)

According to their physical–chemical properties (e.g. molecular size, hydrophobicity and charge), the 20 amino acids (AAs) are further divided into five categories. The five categories are aliphatic group c1: GAVLMI, aromatic group c2: FYW, positively charged groups c3: KRH, negatively charged group c4: DE, and uncharged group c5: STCPNQ. A protein sequence of length L, can be coded as follows.

The grouped amino acid composition (GAAC) [37, 52] descriptor can be defined as:

$$f(c) = \frac{N(c)}{L}; \ y \in \{c1, c2, c3, c4, c5\}; \tag{1}$$

$$N(c_i) = \sum N(i), \quad i \in c \tag{2}$$

where $N(c)$ represents the number of AAs in group c, and $N(i)$ is the number of AAs of type i.

The grouped dipeptide composition (GDPC) [37, 52] encoding is also a variation of the dipeptide composition descriptor. It is defined as:

$$f(x, y) = \frac{N_{xy}}{L - 1}, \quad x, y \in \{c1, c2, c3, c4, c5\}; \tag{3}$$

where N_{xy} is the number of dipeptides represented by AA type groups x and y.

The grouped tripeptide composition (GTPC) [37, 52] encoding is a variation of the tripeptide composition descriptor. It is defined as:

$$f(x, y, z) = \frac{N_{xyz}}{L - 2} \ \ x, y, z \in \{c1, c2, c3, c4, c5\} \tag{4}$$

where N_{xyz} refers to the number of tripeptides represented by AA type groups x, y and z.

GAAPC is a combination of GAAC, GDPC and GTPC. This method will produce 155D feature vectors.

Adaptive skip dipeptide composition (ASDC)

The ASDC descriptor was first presented by Wei et al. [53]. This method is another variant dipeptide composition that considers not only the relevant information between adjacent residues, but also that of intervening residues [54, 55]. It is defined as:

$$ASDC = (FV_1; FV_2; \ldots; FV_{400}) \tag{5}$$

$$FV_i = \frac{\sum_{k=1}^{N-1} f_i^k}{\sum_{i=1}^{400} \sum_{k=1}^{N-1} f_i^k} \tag{6}$$

where FVi represents the occurrence frequency of all possible dipeptides with $\leq N-1$ intervening amino acids. In the ASDC method, the sequence can be easily converted to a 400-dimensional vector.

Pseudo-amino acid composition (PAAC)

The PAAC descriptor is a very effective feature extraction method and is widely used in protein attribute prediction, drug development and studies on drug target areas [56]. The sequence order correlation factors in PAAC incorporate the sequence-order information to some extent. Additional details of the PAAC features are described in the References [56–59]. We used the default parameters in iLearn to obtained a 22-dimensional feature vector.

Classifiers

Six widely used classifiers were investigated to search for the most suitable machine learning algorithm, including random forest (RF), support vector machine (SVM), adaboost (AB), logistic regression (LR), bagging and gradient boosting machine (GBM). These efficient classification models in the scikit-learn package [60] were applied for models implementation and feature importance analysis. The hyper-parameters were optimized using grid search, and the search range was presented in Additional file 1: Table S1.

Feature selection

The features extracted from a sequence in machine learning modeling always contain noise. To improve the feature representation ability, feature selection strategies are often adopted to solve the problems of redundant information and overfitting. Various approaches have been developed to analyze the features, such as analysis of variance (ANOVA) [61–65], minimal redundancy-maximal relevance (MRMR) [66–68] and MRMD [69–72]. These methods have been widely used in the field of RNA, DNA and protein prediction. In this work, MRMD was used to select optimal features for model training. The MRMD feature selection method is mainly determined by two parts [73]. The first part is the correlation between the feature and target class vector calculated by the Pearson correlation coefficient. The second part is the redundancy between features determined by three distance formulas (i.e., Euclidean distance, cosine distance and the Tanimoto coefficient). The larger the Pearson correlation coefficient is, the closer the relationship between the feature and the class label, and the larger the distance is, the lower the redundancy between the features. Finally, MRMD selects a subset of features that are strongly correlated with the class label and have low redundancy between features. We ranked the original features based

Jiao *et al. J Transl Med*   (2021) 19:449

Page 5 of 11

on the MRMD feature sorting algorithm and then applied the IFS strategy to search for the optimal feature subset.

Unbalanced strategy

Data imbalance has been encountered in multiple areas, such as bioinformatics, drug discovery, and disease diagnosis, and has been considered one of the top ten problems in pattern recognition and data mining [74–80]. Fortunately, several approaches have been specifically proposed by researchers to handle such datasets. The data level strategy is a direct way to balance the dataset by increasing/deleting the number of samples in the minority (majority) class. It can be divided into three categories, namely the over-sampling, under-sampling and hybrid-sampling methods [81–83]. In this research, we chose the hybrid-sampling method SMOTE-Tomek to balance the training dataset. This approach is a combination of over- and under-sampling methods: synthetic minority over-sampling technique (SMOTE) [84] and Tomek's links (Tomek) [85]. This hybrid-sampling approach can simultaneously avoid the shortcomings of overfitting and loss of key information caused by SMOTE and Tomek, respectively.

Evaluation parameters and strategies

According to previous related studies, there are three commonly used methods to evaluate the models in the field of protein prediction: K-fold CV, independent test and jackknife test. In this study, we used tenfold CV and independent tests to evaluate and optimize the model. For the binary classification, the confusion-matrix-based metrics are usually applied to measure the predictor, including accuracy (ACC), true negative rate (TNR)/ specificity (Sp), true positive rate (TPR)/sensitivity (Sn), and matthew's correlation coefficient (MCC) [86–101]. However, ACC does not perform well with imbalanced datasets, therefore, balanced accuracy (BACC) was used

to measure how accurate is the overall performance of the models in this work. The formulas for these metrics are presented below:

$$\begin{cases} \text{Sn,} \ \ \text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \\ \text{Sp,} \ \ \text{TNR} = \frac{\text{TN}}{\text{TN+FP}} \\ \text{BACC} = \frac{1}{2} \times (\text{TPR} + \text{TNR}) \\ \text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(\text{TP+FP}) \times (\text{TP+FN}) \times (\text{TN+FP}) \times (\text{TN+FN})}} \end{cases} \quad (7)$$

where TP, TN, FP and FN represent true positive samples, true negative samples, false positive samples and false negative samples, respectively. In addition, the area under the receiver operating characteristic (auROC, also called AUC) curve is also employed, which is used to illustrate the prediction performance of the proposed models.

Results and discussion

Performance of individual feature descriptor

First, we studied the performance of four feature representation methods on six widely used machine learning classifiers. The tenfold CV was used to evaluate all models for fair comparison. The corresponding experimental results were summarized in Additional file 1: Table S2 and the BACC scores were shown in Table 2. As shown in Table 2, the random forest algorithm had the highest BACC on the three features of GPSD, ASDC and GAAPC. Although the BACC of RF on PAAC was not the highest, it was only slightly lower than the highest LR. For the performance of the four feature coding methods on the RF classifier, GPSD had the highest BACC of 69.62%, followed by GAAPC and ASDC, which had BACC of 67.71% and 66.88%, respectively, and PAAC had the lowest BACC of 62.45%.

As discussed earlier, in unbalanced prediction tasks, conventional classifiers usually show poor recognition ability on minority classes. To enhance the performance

Table 2 Preliminary results of different feature descriptors using different classifiers

Features	Classifier (BACC%)					
	LR	Bagging	RF	AB	GBM	SVM
GPSD	63.09	68.51	**69.62**	67.51	65.60	65.81
ASDC	58.53	63.13	**66.88**	62.96	64.03	64.51
GAAPC	58.91	66.75	**67.71**	64.44	67.56	61.75
PAAC	66.01	61.79	62.45	62.68	62.34	**67.91**
GPSD_B[a]	67.12	75.23	**79.62**	73.54	76.91	76.13
ASDC_B[a]	69.50	72.77	**79.19**	71.90	77.02	77.89
GAAPC_B[a]	66.33	75.51	**79.14**	73.02	78.34	74.94
PAAC_B[a]	66.63	71.77	**77.46**	69.91	74.29	73.09

The best performance value is highlighted in bold for clarification

[a] SMOTE-Tomek technique was applied to balance the data set

Jiao *et al. J Transl Med* (2021) 19:449

Page 6 of 11

of these predictive models, we used integrated resampling technique SMOTE-Tomek to balance the positive and negative samples and the results were Additional file 1: Table S3 and the BACC scores were also presented in Table 2. For a more intuitive comparison, the ROC curves before and after resampling were plotted in Additional file 1: Figure S1−4. The performance of all models was significantly improved after using SMOTE-Tomek except for the PAAC on LR classifier. Obviously, RF classifier again top ranked for its high BACC and AUC values on all balanced data. Similar to the results before resampling, the GPSD on the RF classifier showed the best performance among all the feature descriptors with an BACC score of 79.62% and PAAC performed the worst with BACC of 77.46%.

Performance of hybrid features

By using a combination of various feature types, each can alleviate the others' weaknesses and can integrate more sequence information, which helps predict TTCA. Based on these facts, comprehensive prediction performances of hybrid features were further evaluated. We evaluated all possible 11 combinations of four single descriptors, where the imbalanced samples were handled by the SMOTE-Tomek. Using the 11 hybrid features and 6 classifiers, we re-constructed 66 predictive models and the tenfold CV results on training set were provided in Additional file 1: Table S3. As shown in Table S4, almost all hybrid features performed best on RF except for the combination of ASDC+PAAC. This confirmed once again that RF is the most suitable classifier to distinguish TTCA from non-TTCA. In order to find the best hybrid feature to construct the optimal prediction model, we presented the prediction results of RF classifier in Table 3. It can be clearly seen from Table 3 that all hybrid features perform better than the individual features except for ASDC+PAAC combination. There seems no distinct regularity between the combination manner and the performance of corresponding model. The GPSD+GAAPC+PAAC combination yield the best prediction capability among all features in four metrics, with BACC of 83.03%, Sn of 88.69%, Sp of 77.38% and MCC of 0.665. When further integrating ASDC, the overall performance of the model drops sharply, which may be caused by the redundant features introduced by ASDC. Combining the results in Table 2, we can conclude that those models containing GPSD information GPSD are better than those without GPSD information. This indicates that GPSD descriptor is more predictive and discriminative than the others for TTCA prediction. Altogether, the GPSD+GAAPC+PAAC combination outperforms all the features (including individual

Table 3 The classification results of different hybrid features, (1): GPSD, (2): ASDC, (3): GAAPC, (4): PAAC

feature	BACC (%)	AUC	Sn (%)	Sp (%)	MCC
(1) + (2)	80.67	0.875	87.87	73.48	0.620
(1) + (3)	81.26	0.880	86.68	75.85	0.629
(1) + (4)	80.54	0.870	86.20	74.89	0.615
(2) + (3)	81.05	0.863	87.89	74.22	0.627
(2) + (4)	79.01	0.870	83.30	74.73	0.582
(3) + (4)	79.91	0.860	84.28	75.55	0.601
(1) + (2) + (3)	81.12	**0.883**	88.54	73.71	0.629
(1) + (2) + (4)	80.11	0.876	86.74	73.48	0.608
(1) + (3) + (4)	**83.03**	0.882	**88.69**	**77.38**	**0.665**
(2) + (3) + (4)	80.79	0.874	86.24	75.33	0.619
(1) + (2) + (3) + (4)	80.97	0.878	86.71	75.23	0.623

The best performance value is highlighted in bold for clarification

features and hybrid features), and was selected for the next feature analysis experiment.

Performance of optimal feature subset

To determine the optimal feature subset, we first sorted the original 365-dimensional hybrid feature (i.e. GPSD+GAAPC+PAAC obtained in "Performance of hybrid features" section) according to their importance measured by the MRMD algorithm. In the second step, the IFS strategy was applied to further determine the feature vector space for the RF classifier. A total of 365 RF models were trained on 365 feature subsets with 1, 2, 3…, 365 features. The five metrics mentioned above were used to evaluate the models. As shown in Fig. 2A, the tenfold CV BACC scores increased sharply as features were added when the dimension of the feature was less than 60, and then approached slowly fluctuating rising plateaus. When feature dimensions reached 263, the model achieved maximum tenfold CV BACC of 83.71% (Detailed results were presented in Table 4). To build a model with good robustness and generalization, the top 263-dimensional feature subset was selected as the final optimal feature space (named F263).

Moreover, the extensively used data visualization method t-distributed stochastic neighbor embedding (t-SNE) [102]was utilized to validate the effective representation ability of the optimal feature set. We compared our optimal feature F263 with the two best performing individual feature descriptors (GPSD and GAAPC). The t-SNE were calculated for TTCA and non-TTCA of the three compared feature vectors and were plotted in Fig. 2B−D. For the original individual feature descriptors GPSD, the positive samples were randomly distributed in the feature space, while a small number of negative samples were concentrated in the upper and lower positions
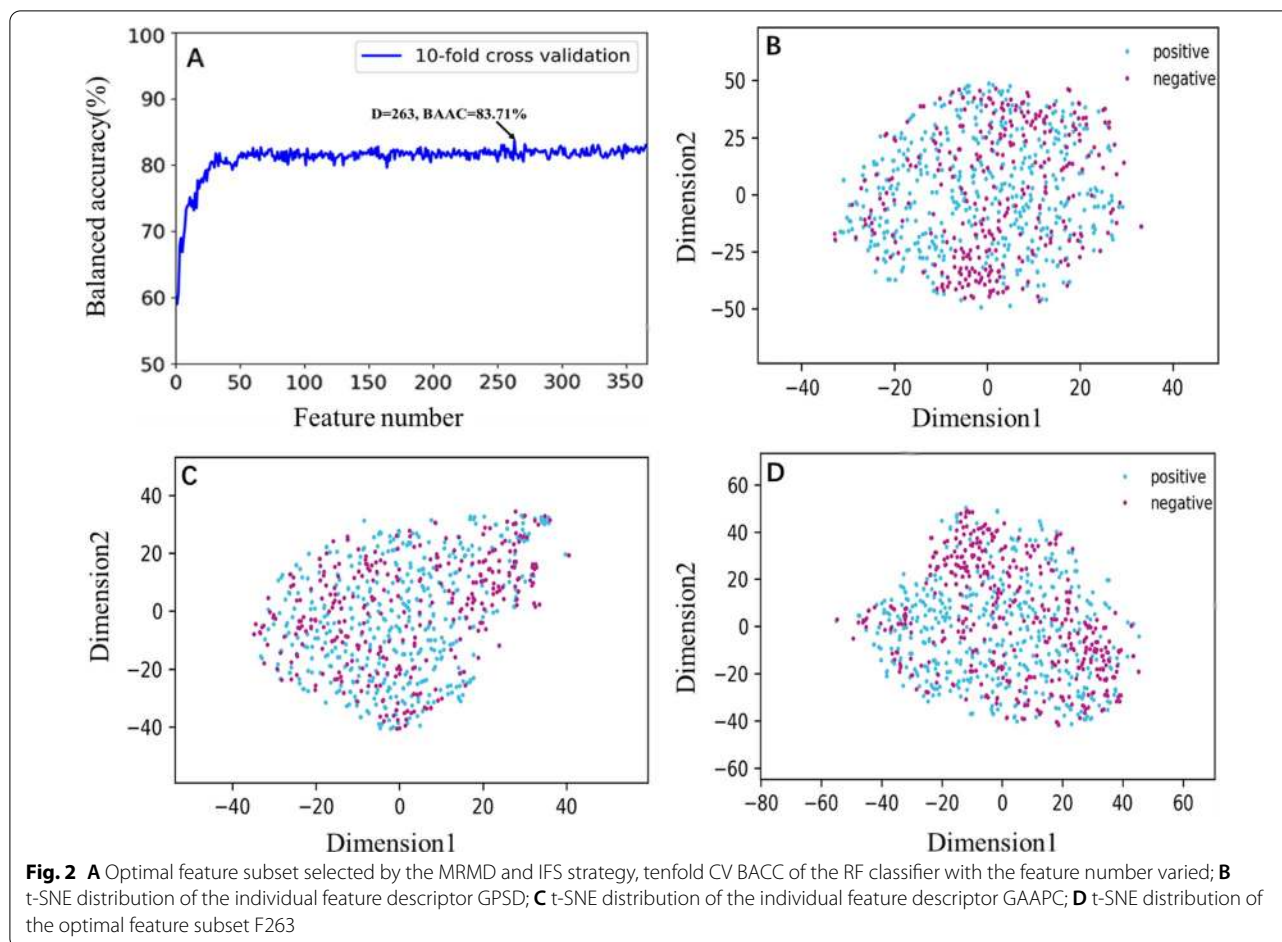
Fig. 2 **A** Optimal feature subset selected by the MRMD and IFS strategy, tenfold CV BACC of the RF classifier with the feature number varied; **B** t-SNE distribution of the individual feature descriptor GPSD; **C** t-SNE distribution of the individual feature descriptor GAAPC; **D** t-SNE distribution of the optimal feature subset F263

Table 4 Comparison of iTTCA-RF and state-of-the-art predictors

Tools	Tenfold CV					Independent test				
	BACC (%)	AUC	Sn (%)	Sp (%)	MCC	BACC (%)	AUC	Sn (%)	Sp (%)	MCC
TTAgP1.0	63.68	0.838	70.85	56.50	**0.838**	68.68	0.747	78.69	58.67	0.379
iTTCA-Hybrid	78.83	0.840	85.53	72.13	0.588	70.73	**0.783**	82.79	58.67	0.428
iTTCA-RF	**83.71**	**0.894**	**88.69**	**78.73**	0.678	**73.14**	0.780	**83.61**	**62.67**	**0.474**

The best performance value is highlighted in bold for clarification

of the graph. As for GAAPC, most of positive and negative samples were randomly distributed and overlapping. For these two distribution maps, most samples overlap, and it was difficult to fit a boundary that distinguished the two types of samples. However, shown in Fig. 2D, although the distribution of positive samples and negative samples in the optimal F263 feature space still overlapped somewhat, it was simpler and clearer to find the dividing line that could distinguish most negative samples from positive samples. This indicates that using the 263-dimensional feature subset obtained of hybrid features by MRMD is easier to identify TTCA and

non-TTCA samples than when using the original individual feature descriptors. Therefore, the tenfold CV results of iTTCA-RF were improved.

Comparison with reported tools

Two classifiers to discriminate TTCA and non-TTCA have been published: TTAgP1.0 and iTTCA-Hybrid. Table 4 summarizes the tenfold CV and independent test scores of the three predictors. The results of TTAgP1.0 were from TTAgP1.0-MODI established by Charoenkwan et al. using the same method on the new dataset. Thus, all three tools were compared on the same training

Jiao *et al. J Transl Med*   (2021) 19:449

Page 8 of 11

and testing dataset. Since almost all the results metrics of iTTCA-Hybrid were better than TTAgP1.0, we mainly compared iTTCA-RF with iTTCA-Hybrid.

Figure 3A also visually demonstrates the comparison of evaluation metrics, and ROC curves were drawn (Fig. 3B) to depict the prediction efficiency. As shown in Fig. 3, the tenfold CV of iTTCA-RF scores was higher than that of the iTTCA-Hybrid scores in almost all metrics. The BACC, AUC, Sn, Sp and MCC of our model on the training set were 4.9%, 5.4%, 3.2%, 6.6% and 9.0% higher than those of iTTCA-Hybrid, respectively. In terms of independent test scores, BACC, Sn, Sp and MCC outperformed iTTCA-Hybrid with improvement of 2.4%, 0.8%, 4.0% and 4.6%, respectively. These results indicate that the prediction capacity on negative samples of iTTCA-RF was greatly improved compared with the other predictors. Overall, iTTCA-RF significantly outperformed the other latest predictors, indicating that it can distinguish true TTCA from non-TTCA more accurately than existing tools. Although the developed predictor showed good performance, there is still much room for improvement, especially in terms of the predictive ability on negative samples.

Web server implementation

For convenience, a user-friendly online server called iTTCA-RF was developed, which can be accessed freely at http://lab.malab.cn/~acy/iTTCA. Users can use the web-server to identify whether their protein sequences (in FASTA format) are TTCA or non-TTCA. The first step is to enter or paste the FASTA format protein sequences in the left blank box and then click the Submit button. The identification results will be displayed in the box on the right. If starting a new task, the user needs to click the Clear button or the Resubmit button to clear the input box. The Submit button will be reactivated, and the user will be allowed to input new query protein sequences. The homepage also provides links to download relevant data and contact the author.

Conclusion

Accurate identification of TTCA will greatly promote cancer vaccine research and development. In this study, we constructed a new computational TTCA identifier named iTTCA-RF using the hybrid features of GPSD, GAAPC and PAAC. Combining the feature selection technique MRMD followed by IFS theory, the top 263 important features were chosen to build the best performance predictor. Here, the imbalance problem was addressed using the resampling method SMOTE-Tomek. iTTCA-RF achieves the best CV evaluation BACC value of 83.71%, which is 4.9% higher than the corresponding value of the previously reported best predictor. The independent test BACC score was 73.14%, an improvement of 2.4%, and associated Sp and MCC values were also increased by 4.0% and 4.6%, respectively. Meanwhile, a user-friendly web-server was also established. It is expected that iTTCA-RF will be a robust, reliable, and useful computational tool for tumor T cell antigen identification. Although our proposed model is superior to other published predictors, the model requires further development, especially the ability to identify negative samples. Future work will focus on exploring deep
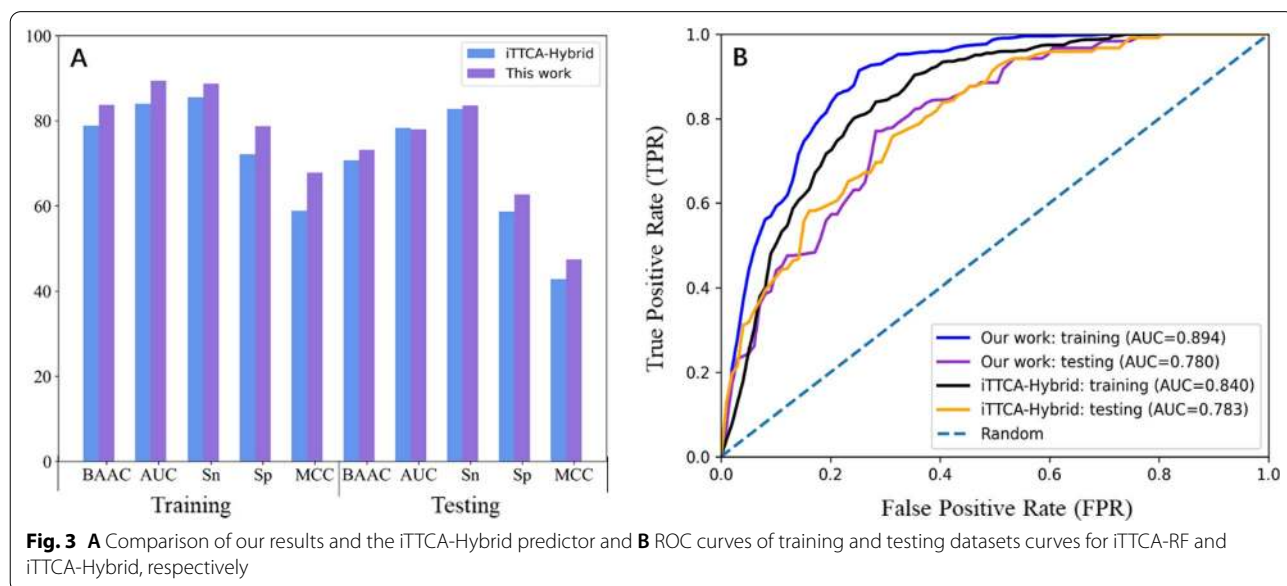

Fig. 3 **A** Comparison of our results and the iTTCA-Hybrid predictor and **B** ROC curves of training and testing datasets curves for iTTCA-RF and iTTCA-Hybrid, respectively

Jiao *et al. J Transl Med*   (2021) 19:449

Page 9 of 11

learning [103, 104] or more effective feature representation or computational intelligence strategies to improve the model's performance.

Abbreviations
IARC: International Agency for Research on Cancer; MHC: Major histocompatibility complex; TTCA: Tumor T cell antigens; AAC: Amino acid composition; RF: Random Forest; PAAC: Pseudo amino acid composition; CTDD: Amino acid property distribution; GPSD: Global protein sequence descriptors; GAAPC: Grouped amino acid and peptide composition; ASDC: Adaptive skip dipeptide composition; MRMD: Maximum relevance maximum distance; IFS: Incremental feature selection; CV: Cross-validation; BACC: Balanced accuracy; AAs: Amino acids; GAAC: Grouped amino acid composition; GDPC: Grouped dipeptide composition (GDPC); GTPC: Grouped tripeptide composition; SVM: Support vector machine; AB: Adaboost; LR: Logistic regression; GBM: Gradient boosting machine; ANOVA: Analysis of variance; SMOTE: Synthetic minority over-sampling technique; ACC: Accuracy; TNR: True negative rate; TPR: True positive rate; Sp: Specificity; Sn: Sensitivity; MCC: Matthew's correlation coefficient; AUC: Area under the receiver operating characteristic curve.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12967-021-03084-x.

Additional file 1: Table S1. Hyperparameters search range for Random Forest classifier. **Table S2.** Results of single feature descriptors without SMOTE-Tomek over 10-fold CV. **Table S3.** Results of single feature descriptor with SMOTE-Tomek over 10-fold CV. **Table S4.** Results of hybrid feature descriptor with SMOTE-Tomek over 10-fold CV. **Figure S1.** 10-fold CV ROC curves for GPSD descriptor on six classifiers. **Figure S2.** 10-fold CV ROC curves for ASDC descriptor on six classifiers. **Figure S3.** 10-fold CV ROC curves for GAAPC descriptor on six classifiers. **Figure S4.** 10-fold CV ROC curves for PAAC descriptor on six classifiers.

Acknowledgements
Not applicable.

Authors' contributions
QZ, HG and LS contributed to the conception of the study; SJ performed the experiment; QZ and HG analyzed the results; SJ and HG wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

Availability of data and materials
Publicly available datasets were analyzed in this study. This data can be found here: http://lab.malab.cn/~acy/iTTCA.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no potential conflicts of interest.

Author details
[1] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China. [2] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China. [3] Department of Oncology, General Hospital of Heilongjiang Province Land Reclamation Bureau, Harbin, China. [4] Department of Spine Surgery, Changzheng Hospital, Naval Medical University, Shanghai, China.

References
1. Zhang ZM, et al. Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine-learning method. Front Cell Dev Biol. 2020;8:582864.
2. Cheng L, et al. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics. 2018;34(11):1953–6.
3. Burugu S, Dancsok AR, Nielsen TO. Emerging targets in cancer immunotherapy. Semin Cancer Biol. 2018;52:39–52.
4. Dong Y-M, et al. ESDA: an improved approach to accurately identify human snoRNAs for precision cancer therapy. Curr Bioinform. 2020;15(1):34–40.
5. Yu L, et al. Predicting therapeutic drugs for hepatocellular carcinoma based on tissue-specific pathways. PLoS Comput Biol. 2021;17(2):e1008696.
6. Behl T, et al. Gene therapy in the management of Parkinson's disease: potential of gdnf as a promising therapeutic strategy. Curr Gene Ther. 2020;20(3):207–22.
7. Couzin-Frankel J. Cancer immunotherapy. Science. 2013;342(6165):1432–3.
8. Li Z, et al. Research on gastric cancer's drug-resistant gene regulatory network model. Curr Bioinform. 2020;15(3):225–34.
9. Ding Y, Tang J, Guo F. Identification of drug-target interactions via dual laplacian regularized least squares with multiple kernel fusion. Knowl Based Syst. 2020;204:106254.
10. Ding Y, Tang J, Guo F. Identification of drug-target interactions via fuzzy bipartite local model. Neural Comput Appl. 2020;23:10303–19.
11. Ding Y, Tang J, Guo F. Identification of drug-target interactions via multiple information integration. Inf Sci. 2017;418:546–60.
12. Zhang G, et al. TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes. BMC Bioinform. 2021;22:1–8.
13. Zhao X, et al. Predicting drug side effects with compact integration of heterogeneous networks. Curr Bioinform. 2019;14(8):709–20.
14. Ding Y, Tang J, Guo F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing. 2019;325:211–24.
15. Shang Y, et al. Prediction of drug-target interactions based on multi-layer network representation learning. Neurocomputing. 2021;434:80–9.
16. Aranda F, et al. Trial watch peptide vaccines in cancer therapy. Oncoimmunology. 2013;2(12):e26621.
17. Liu Y, et al. A review on the methods of peptide-MHC binding prediction. Curr Bioinform. 2020;15(8):878–88.
18. Wang P, et al. Comprehensive analysis of TCR repertoire in COVID-19 using single cell sequencing. Genomics. 2020;113(2):456–62.
19. Ren X, et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. Cell. 2021;184(7):1895-1913.e19.
20. Liu K, Chen W. iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. Bioinformatics. 2020;36(11):3336–42.
21. Ao C, Yu L, Zou Q. Prediction of bio-sequence modifications and the associations with diseases. Brief Funct Genomics. 2021;20(1):1–18.
22. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. 2019;47(20):e127.
23. Zulfiqar H, et al. Screening of prospective plant compounds as H1R and CL1R inhibitors and its antiallergic efficacy through molecular docking approach. Comput Math Methods Med. 2021;2021:6683407.

Jiao *et al. J Transl Med*   (2021) 19:449

Page 10 of 11

24. Yang H, et al. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. Inf Fus. 2021;75:140–9.
25. Yu L, Shi Y, Zou Q, Wang S, Zheng L, Gao L. Exploring drug treatment patterns based on the action of drug and multilayer network model. Int J Mol Sci. 2020;21(14):5014.
26. Fu X, et al. StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. Bioinformatics. 2020;36(10):3028–34.
27. Zeng X, et al. Target identification among known drugs by deep learning from heterogeneous networks. Chem Sci. 2020;11(7):1775–97.
28. Zeng X, et al. Prediction and validation of disease genes using HeteSim scores. IEEE/ACM Trans Comput Biol Bioinf. 2017;14(3):687–95.
29. Cheng L, et al. MetSigDis: a manually curated resource for the metabolic signatures of diseases. Brief Bioinform. 2019;20(1):203–9.
30. Hu Y, et al. rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues. BMC Med. 2021;19(1):11.
31. Beltran Lissabet JF, Herrera Belen L, Farias JG. TTAgP 10: a computational tool for the specific prediction of tumor T cell antigens. Comput Biol Chem. 2019;83:107103.
32. Ao C, et al. Prediction of antioxidant proteins using hybrid feature representation method and random forest. Genomics. 2020;112(6):4666–74.
33. Charoenkwan P, et al. iTTCA-Hybrid: Improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. Anal Biochem. 2020;599:113747.
34. Olsen LR, et al. TANTIGEN: a comprehensive database of tumor T cell antigens. Cancer Immunol Immunother. 2017;66(6):731–5.
35. Vita R, et al. The immune epitope database (IEDB): 2018 update. Nucleic Acids Res. 2019;47(D1):D339–43.
36. Muhammod R, et al. PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. Bioinformatics. 2019;35(19):3831–3.
37. Chen Z, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. 2020;21(3):1047–57.
38. Wang H, et al. Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence Criterion. Neurocomputing. 2020;383:257–69.
39. Li J, et al. DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides. IEEE J Biomed Health Inform. 2020;24(10):3012–9.
40. Shen Y, Tang J, Guo F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. J Theor Biol. 2019;462:230–9.
41. Shen Y, et al. Critical evaluation of web-based prediction tools for human protein subcellular localization. Brief Bioinform. 2019;21:1628–40.
42. Tang Y-J, Pang Y-H, Liu B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. Bioinformaitcs. 2020;36(21):5177–86.
43. Shao J, Yan K, Liu B. FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbaa144.
44. Cai L, et al. ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation. Brief Bioinform. 2020;22:bbaa367.
45. Jin S, et al. Application of deep learning methods in biological networks. Brief Bioinform. 2020;22(2):1902–17.
46. Zhao T, et al. DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. Bioinformatics. 2020;36:4466–72.
47. Dubchak I, et al. Prediction of protein-folding class using global description of amino-acid-sequence. Proc Natl Acad Sci USA. 1995;92(19):8700–4.
48. Zou Q, et al. An approach for identifying cytokines based on a novel ensemble classifier. Biomed Res Int. 2013. https://doi.org/10.1155/2013/686090.
49. Li Y, Niu M, Zou Q. ELM-MHC: an improved MHC identification method with extreme learning machine algorithm. J Proteome Res. 2019;18(3):1392–401.
50. Xuan JJ, et al. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res. 2018;46:D327–34.
51. Lin C-W, et al. Kaempferol reduces matrix metalloproteinase-2 expression by down-regulating ERK1/2 and the activator protein-1 signaling pathways in oral cancer cells. PLoS ONE. 2013;8(11):e80883.
52. Chen Z, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;34(14):2499–502.
53. Wei L, Tang J, Zou Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. BMC Genom. 2017;18:1.
54. Wei L, et al. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018;34(23):4007–16.
55. Zhang D, et al. iBLP: an XGBoost-based predictor for identifying bioluminescent proteins. Comput Math Methods Med. 2021;2021:6664362.
56. Xu L, et al. A novel hybrid sequence-based model for identifying anticancer peptides. Genes. 2018;9(3):158.
57. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol. 2011;273(1):236–47.
58. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005;21(1):10–9.
59. Liu B, Zhu Y, Yan K. Fold-LTR-TCP: protein fold recognition based on triadic closure principle. Brief Bioinform. 2020;21(6):2185–93.
60. Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
61. Blanca MJ, et al. Non-normal data: is ANOVA still a valid option? Psicothema. 2017;29(4):552–7.
62. Tang H, Chen W, Lin H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Mol BioSyst. 2016;12(4):1269–75.
63. Jung Y, Zhang H, Hu J. Transformed low-rank ANOVA models for high-dimensional variable selection. Stat Methods Med Res. 2019;28(4):1230–46.
64. Tan JX, et al. Identification of hormone binding proteins based on machine learning methods. Math Biosci Eng. 2019;16(4):2466–80.
65. Han X, et al. SubtypeDrug: a software package for prioritization of candidate cancer subtype-specific drugs. Bioinformatics. 2021. https://doi.org/10.1093/bioinformatics/btab011.
66. Ju Z, Wang S-Y. iLys-Khib: identify lysine 2-Hydroxyisobutyrylation sites using mRMR feature selection and fuzzy SVM algorithm. Chemom Intell Lab Syst. 2019;191:96–102.
67. Mostafa SS, Morgado-Dias F, Ravelo-Garcia AG. Comparison of SFS and mRMR for oximetry feature selection in obstructive sleep apnea detection. Neural Comput Appl. 2020;32(20):15711–31.
68. Wang J, Zhang D, Li J. PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection. BMC Syst Biol. 2013;7:1–9.
69. Meng C, et al. CWLy-pred: a novel cell wall lytic enzyme identifier based on an improved MRMD feature selection method. Genomics. 2020;112(6):4715–21.
70. Tao Z, et al. A method for identifying vesicle transport proteins based on LibSVM and MRMD. Comput Math Methods Med. 2020. https://doi.org/10.1155/2020/8926750.
71. He S, et al. MRMD2.0: a python tool for machine learning with feature ranking and reduction. Curr Bioinform. 2020;15(10):1213–21.
72. Lu XX, Zhao SZ. Gene-based therapeutic tools in the treatment of cornea disease. Curr Gene Ther. 2019;19(1):7–19.
73. Zou Q, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing. 2016;173:346–54.
74. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18:559–63.
75. Yang X-F, et al. Predicting LncRNA subcellular localization using unbalanced pseudo-k nucleotide compositions. Curr Bioinform. 2020;15(6):554–62.
76. Hasan MAM, et al. Citrullination site prediction by incorporating sequence coupled effects into PseAAC and resolving data imbalance issue. Curr Bioinform. 2020;15(3):235–45.

Jiao *et al. J Transl Med*   (2021) 19:449

Page 11 of 11

77. Chao L, Wei L, Zou Q. SecProMTB: a SVM-based classifier for secretory proteins of *Mycobacterium tuberculosis* with imbalanced data set. Proteomics. 2019;19:e1900007.

78. Yu L, et al. Prediction of drug response in multilayer networks based on fusion of multiomics data. Methods (San Diego, Calif). 2020;192:85–92.

79. Zeng X, et al. A comprehensive overview and evaluation of circular RNA detection tools. Plos Comput Biol. 2017;13(6):e1005420.

80. Zeng X, et al. Prediction of potential disease-associated micro-RNAs using structural perturbation method. Bioinformatics. 2018;34(14):2425–32.

81. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv. 2019;52(4):1–36.

82. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on IM balanced domains. ACM Comput Surv. 2016;49(2):1–50.

83. Zou Q, et al. Finding the best classification threshold in imbalanced classification. Big Data Res. 2016;5:2–8.

84. Chawla NV, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.

85. Tomek I. Two modifications of CNN. IEEE Trans Syst Man Cybern. 1976;SMC6(11):769–72.

86. Wang H, Tang J, Ding Y, Guo F. Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hyper-graph-regular terms on center kernel alignment. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbaa409.

87. Li J, Pu Y, Tang J, Zou Q, Guo F. DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. Brief Bioinform. 2020;22:bbaa59.

88. Hong Z, et al. Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics. 2020;36(4):1037–43.

89. Jin Q, et al. DUNet: a deformable network for retinal vessel segmentation. Knowl-Based Syst. 2019;178:149–62.

90. Su R, et al. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief Bioinform. 2020;21(2):408–20.

91. Wei L, Chen H, Su R. M6APred-EL: a sequence-based predictor for identifying n6-methyladenosine sites using ensemble learning. Mol Ther Nucleic Acids. 2018;12:635–44.

92. Wei L, et al. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. Brief Bioinform. 2020;22:bbaa275.

93. Wei L, et al. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief Bioinform. 2020;21(1):106–19.

94. Wei L, et al. Improved and promising identification of human Micro-RNAs by incorporating a high-quality negative set. IEEE/ACM Trans Comput Biol Bioinf. 2014;11(1):192–201.

95. Wei L, et al. A novel hierarchical selective ensemble classifier with bioinformatics application. Artif Intell Med. 2017;83:82–90.

96. Wei L, et al. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif Intell Med. 2017;83:67–74.

97. Shao J, Liu B. ProtFold-DFG: protein fold recognition by combining directed fusion graph and PageRank algorithm. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbaa192.

98. Jiang Q, et al. Predicting microRNA-disease associations based on support vector machine. Int J Data Min Bioinform. 2013;8(3):282–93.

99. Yu L, Xu F, Gao L. Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression. Front Bioeng Biotechnol. 2020;8:8.

100. Zeng X, et al. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019;35(24):5191–8.

101. Hong Z, et al. Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics. 2019;36(4):1037–43.

102. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.

103. Lv H, et al. Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. Brief Bioinform. 2020;22:bbaa255.

104. Dao FY, et al. DeepYY1: a deep learning approach to identify YY1-mediated chromatin loops. Brief Bioinform. 2020;22:bbaa356.

Publisher's Note