RESEARCH ARTICLE

A novel feature ranking method for prediction of cancer stages using proteomics data

Ehsan Saghapour, Saeed Kermani*, Mohammadreza Sehhati

Department of Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

* kermani@med.mui.ac.ir

Abstract

Proteomic analysis of cancers' stages has provided new opportunities for the development of novel, highly sensitive diagnostic tools which helps early detection of cancer. This paper introduces a new feature ranking approach called FRMT. FRMT is based on the Technique for Order of Preference by Similarity to Ideal Solution method (TOPSIS) which select the most discriminative proteins from proteomics data for cancer staging. In this approach, outcomes of 10 feature selection techniques were combined by TOPSIS method, to select the final discriminative proteins from seven different proteomic databases of protein expression profiles. In the proposed workflow, feature selection methods and protein expressions have been considered as criteria and alternatives in TOPSIS, respectively. The proposed method is tested on seven various classifier models in a 10-fold cross validation procedure that repeated 30 times on the seven cancer datasets. The obtained results proved the higher stability and superior classification performance of method in comparison with other methods, and it is less sensitive to the applied classifier. Moreover, the final introduced proteins are informative and have the potential for application in the real medical practice.

Introduction

Cancer has always been one of the most fundamental health problems of the human society. Every year, between 100 and 350 out of every 100,000 people die due to cancer in the worldwide [1–4]. Understanding the nature of cancer, which caused by the malfunction of the mechanisms that regulate growth and cell division, has always been a topic of interest to researchers. The development of molecular biology in recent decades enhanced understanding of complex interactions of the genetic variants, transcription and translation [5]. Proteomic studies can play a critical role in prevention, early detection and treatment of cancer. Given that proteomic studies can help identify cancer biomarkers, it might cause early detection and treatment of cancer [6, 7].

The robustness of microarray-derived cancer biomarkers that have been identified by using gene expression profiles is very poor [8, 9]. Thus, the evaluation of tumor cells at protein expression levels, which are more robust than gene expression level, is necessary to explain

Citation: Saghapour E, Kermani S, Sehhati M (2017) A novel feature ranking method for prediction of cancer stages using proteomics data. PLoS ONE 12(9): e0184203. https://doi.org/10.1371/journal.pone.0184203

Editor: Yong Deng, Southwest University, CHINA

Received: April 27, 2017

Accepted: August 15, 2017

Published: September 21, 2017

Data Availability Statement: All relevant data are available at the following pages: https://github.com/E-Saghapour/FRMT, https://confluence.broadinstitute.org/display/GDAC/Home, and http://tcpaportal.org/tcpa/download.html.

Funding: This research was financially supported by the Isfahan University of Medical Sciences, Isfahan, Iran (Grant no.395775).

Competing interests: The authors have declared that no competing interests exist.

causes of tumor proliferation. And it will help us to find potential drug targets and to illustrate off-target effects in cancer medicine [10].

Zhang et al. [10] utilized the protein expression profiles for classifying ten types of cancers. They applied minimum redundancy maximum relevancy (mRMR) and incremental feature selection (IFS) methods for selecting 23 out of 187 proteins on the protein array, which used as the inputs of sequential minimal optimization (SMO) classifier. Sonntag et al. [11] have introduced a novel biomarker selection workflow to extract four discriminative biomarkers from reverse phase protein array (RPPA) data on luminal breast cancer.

Kaddi and Wang [12] employed three different approaches for feature selection (two filter and one wrapper methods) and six methods for classification (four individual binary and two ensemble classification methods) to predict early stage of cancer in Head and Neck Squamous Cell Carcinoma using proteomic and transcriptomic data.

Stafford et al. [13] randomly generated two libraries, each of them contained approximately 10000 peptide sequences, then they used ANOVA and t-test for feature selection and the linear discriminant analysis (LDA), naive bayes (NB) and support vector machine (SVM) for classification. Numerous studies have been reported for identification of biomarkers that influence in the early detection of ovarian cancer [14].

Nguyen et al. [15, 16] presented a novel feature selection method by integrating the five filter-based feature selection approaches (i.e., t-test, ROC, Wilcoxon, Entropy, and SNR) through an analytic hierarchy process (AHP). AHP, which is a multi-criteria decision analysis method, is used for classification of normal and tumor tissue by means of different classifier algorithms (i.e. Interval type-2 Fuzzy Logic (FL) [17], Hidden markov model (HMM) [18], k-nearest neighbors (kNN) [19], support vector machine (SVM) [20], etc.).

In this research, we proposed a hybrid model for prediction of cancer stages using RPPA data. The novelty of the proposed method relies on the feature ranking using TOPSIS. To improve the stability and accuracy of the final extracted biomarkers, we modified the feature selection workflow and utilized the best classification model among the well-known seven classifiers (i.e. SVM [20], Random Forest (RF) [21], Decision Tree (DT) [22], LDA [23], NB [24], FL [25], and kNN [19]).

As demonstrated by a series of recent publications [26–32], and in agreement with the famous 5-step rule [33], we should comply with the following five step instruction to construct a really useful prediction method for a biomedical system; (1) select or construct a valid benchmark dataset to train and test the predictor model, (2) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) develop or introduce a powerful algorithm to run the prediction, (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor, (5) establish a user-friendly and publicly accessible web-server for the predictor.

The rest of the paper is organized as follows: In Section 2, the utilized database is introduced and a detailed description of the proposed protein ranking and classification methods is presented. In Section 3, the evaluation results of the various protein selection methods in combination with the various kinds of classifiers are described. The related issues of cancer classification are discussed in Section 4. We conclude the paper in Section 5.

Materials and methods

Dataset

Proteomic data, including 2101 patient samples from 7 cancer types were downloaded from The Cancer Proteome Atlas (TCPA) [34]. For each sample, the expressions of 187 proteins

Table 1. Summary of the utilized cancer datasets.

Cancer	# early-stage	# advanced-stage	# Total
READ	60	62	122
HSNE	48	152	200
LUSC	158	35	193
COAD	187	139	326
OV	33	370	403
UCEC	321	83	404
KIRC	263	190	453
Total Number of Samples	1070	1031	2101

READ: Rectum adenocarcinoma, HSNE: Head and Neck sequamous cell carcinoma, LUSC: Lung sequamous cell carcinoma, COAD: Colon adenocarcinoma, OV: Ovarian serous cystadenocarcinoma, UCEC: Uterine Corpus Endometrioid Carcinoma. KIRC: Kideny renal clear cell carcinoma.

https://doi.org/10.1371/journal.pone.0184203.t001

were taken by RPPA. We used RPPA, an antibody-based high-throughput technique, for analyzing concurrent expression levels of hundreds of proteins in a single experiment.

The related pathological information for each patient in the TCPA dataset was downloaded from Broad Institute TCGA (https://confluence.broadinstitute.org/display/GDAC/Dashboard-Stddata and http://tcpaportal.org/tcpa/download.html). Then, we divided the samples into two groups of early stage (stage I and II) and advanced stage (stage III and IV).

In TCPA, the proteins are divided into three groups including "validated", "under evaluation" and "used with caution". In this work, we only used 115 validated labeled proteins per patient to obtain reliable results. See Table 1 for details.

We used the R software and FEAST Toolbox [35] in MATLAB to implement different classification models and feature selection algorithms. All the cleaned data and the algorithms scripts used in this manuscript, can be downloaded from www.github.com/E-Saghapour/FRMT.

Hybrid models

The hybrid model approaches were used by many previous investigators to study various biological or biomedical problems [36–40]. The stratification of cancer can be considered as traditional pattern recognition problems. Data analysis procedure, including feature selection and classification steps, is shown in Fig 1. The explanation of different blocks in Fig 1 is presented in the next subsections.

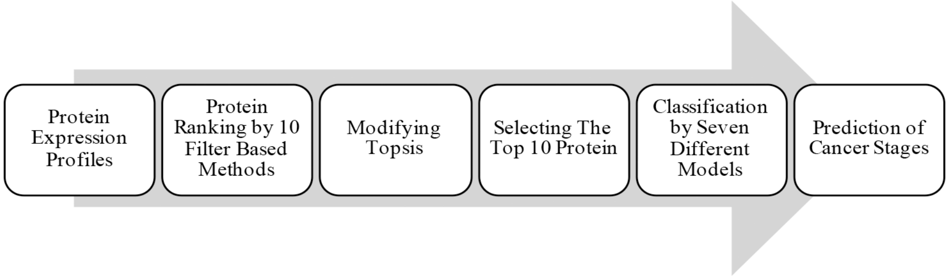

Protein Expression Profiles
Protein Ranking by 10 Filter Based Methods
Modifying Topsis
Selecting The Top 10 Protein
Classification by Seven Different Models
Prediction of Cancer Stages

Fig 1. Schematic of the proposed data analysis procedure. The whole procedure from processing the protein expression profiles to prediction of the cancer stages is illustrated.

https://doi.org/10.1371/journal.pone.0184203.g001

Table 2. Common filter-based feature selection methods.

Criterion	Full name
t-test	Two sample t-test
wrs	Wilcoxon rank sum
mrmr	Max-Relevance Min-Redundancy
mim	Mutual Information Maximization
mifs	Mutual Information Feature Selection
jmi	Joint Mutual Information
disr	Double Input Symmetrical Relevance
cmim	Conditional Mutual Info Maximization
icap	Interaction Capping
cife	Conditional Infomax Feature Extraction

https://doi.org/10.1371/journal.pone.0184203.t002

Feature selection. In a filter feature selection (FFS) method, a criterion function would be used for independently ranking features. Then, the top ranked features, called informative features, would be used in the classification model. Various criterion functions have been introduced and applied to the gene expression profiles that led to different subset of genes with different classification performance. Although the FFS methods produce unstable results in different datasets, but they are robust against overfitting. FFS methods can also be applied to the protein expression profiles for protein ranking, however, they do not take into account protein-protein interactions. In this study, a novel ensemble method is proposed to improve the stability of results obtained by integrating common FFS methods (Table 2). We utilized the TOPSIS method to score the proteins and choose the most informative ones for classification (Fig 2). The TOPSIS method is described in detail in the next section.

TOPSIS method. The TOPSIS was first presented by Hwang and Yoon in 1981 [41]. It is a multi-criteria decision analysis method relied on selecting the option that its geometric distances from the positive ideal solution (PIS) and the negative ideal solution (NIS) are the shortest and longest, respectively.

The workflow of the TOPSIS method contains the following seven steps:

1. Generating an m-by-n evaluation matrix contains m alternatives $A_1, A_2, \ldots, A_m$, each assessed by n local criteria $C_1, C_2, \ldots, C_n$.
2. Normalizing the decision matrix:

$$u_{ij} = \frac{x_{ij}}{\sqrt{\sum_{k=1}^{m} x_{kj}^2}}; \quad i = 1, \ldots, m; \quad j = 1, \ldots., n. \tag{1}$$

Where x_{ij} is the score of alternative A_i with respect to the criterion C_j.

3. Calculating the weighted normalized decision matrix which its values V_{ij} are computed as:

$$V_{ij} = W_i \times u_{ij}; \quad j = 1, 2, \ldots, m; \quad i = 1, 2, \ldots, n.$$

let $W_i = [w_1, w_2, \ldots, w_n]$ be the vector of local criteria weights satisfying $\sum_{i=1}^{n} W_i = 1.$

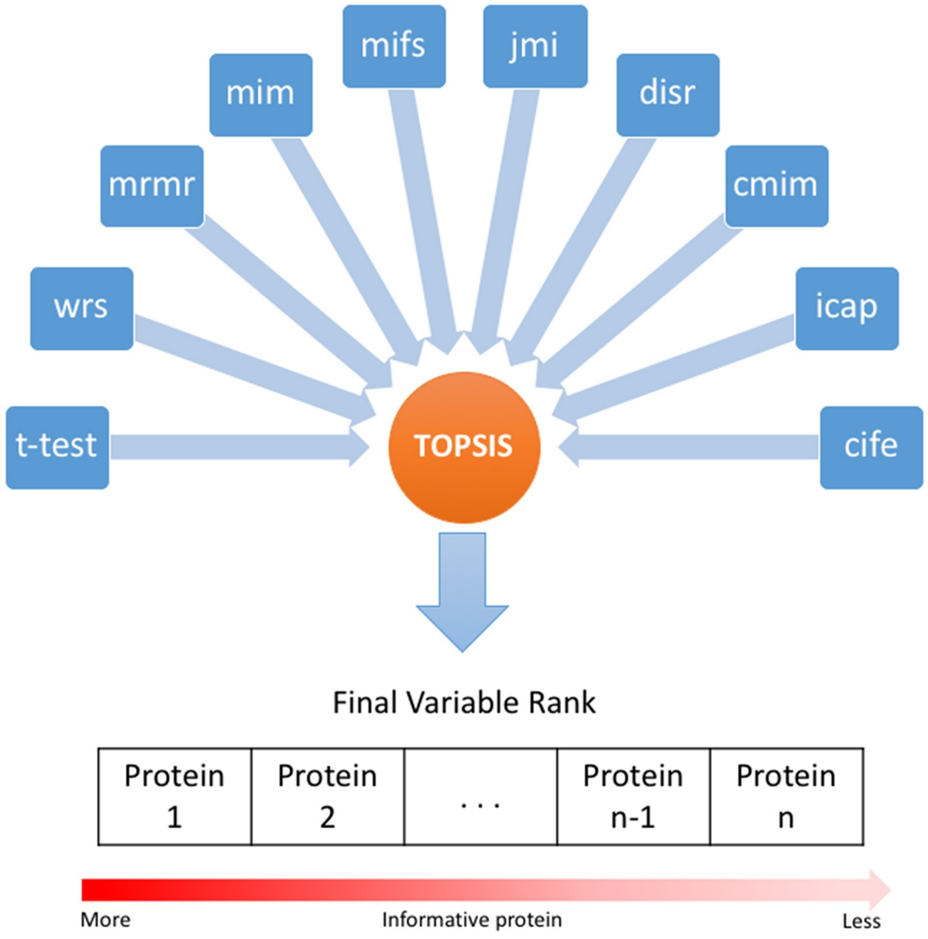


Fig 2. Feature ranking procedure. TOPSIS is used for integration of different FFS methods for proteins ranking.

https://doi.org/10.1371/journal.pone.0184203.g002

4. Determining the positive ideal (A^+) and negative ideal (A^-) solutions as follows:

$$A^+ = \{v_1^+, \ldots, v_n^+\} = \{(\max_i V_{ij}|\ j \in J),\ (\min_i V_{ij}|\ j \in J')\}. \tag{2}$$

$$A^- = \{v_1^-, \ldots, v_n^-\} = \{(\min_i V_{ij}|\ j \in J), (\max_i V_{ij}|\ j \in J')\}. \tag{3}$$

$$J = \{j = 1, 2, 3, \ldots, n | j \ associated \ with \ benefit \ criteria\}. \tag{4}$$

$$J' = \{j = 1, 2, 3, \ldots, n | j \ associated \ with \ cost \ criteria\}. \tag{5}$$

In the proposed method, all criteria are considered as benefit, therefor J' is empty and (2),

(3) will be reduced to (6), (7);

$$A^{+} = \{v_1^{+}, \ldots,\ v_n^{+}\} = \{\max_i V_{ij} | j \in J\}. \tag{6}$$

$$A^{-} = \{v_1^{-}, \ldots,\ v_n^{-}\} = \{\min_i V_{ij} | j \in J\}. \tag{7}$$

5. Measuring the Euclidean Distances between each alternative and both the positive and negative ideal, which are calculated as follows:

$$P_i^{+} = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^{+})^2};\ i = 1, 2, \ldots, m. \tag{8}$$

$$P_i^{-} = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^{-})^2};\ i = 1, 2, \ldots, m. \tag{9}$$

6. Computing the relative closeness to the ideal solution by Eq (10).

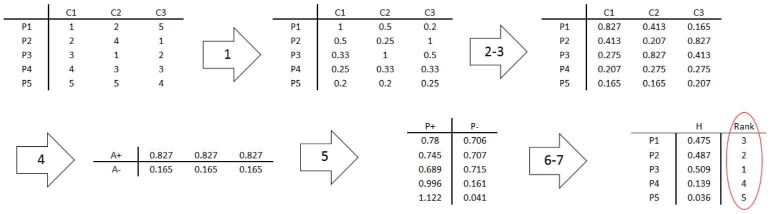
$$H_i = \frac{P_i^{-}}{P_i^{+} + P_i^{-}}\ ;\ i = 1, 2, \ldots, m;\ 0 \leq\ H_i\ \leq 1. \tag{10}$$

7. Ranking alternatives based on the H value of each parameter. $H_i = 1$ indicates the highest rank and $H_i = 0$ indicates the lowest rank.

Fig 3 illustrates the whole procedure of TOPSIS method for a simple example in which we have 5 alternatives (proteins expression), P1-P5, and 3 criteria (methods of feature selection), C1-C3. Moreover, we considered equal weights for all feature selection methods (criteria).

Classification. In this study, we utilized seven models for classification including SVM, RF, DT, LDA, NB, FL, and kNN.

SVM rely on the concept of decision planes that specify decision borders. Classification task performed by building hyperplanes in a multidimensional space that distinct various class labels. The classes that have nonlinear boundaries in the input space employ the kernel function method to map the input space in to a higher dimensional feature space in which linear differentiation may be feasible. The kernel trick computes all training data without using or



	C1	C2	C3
P1	1	2	5
P2	2	4	1
P3	3	1	2
P4	4	3	3
P5	5	5	4

	C1	C2	C3
P1	1	0.5	0.2
P2	0.5	0.25	1
P3	0.33	1	0.5
P4	0.25	0.33	0.33
P5	0.2	0.2	0.25

	C1	C2	C3
P1	0.827	0.413	0.165
P2	0.413	0.207	0.827
P3	0.275	0.827	0.413
P4	0.207	0.275	0.275
P5	0.165	0.165	0.207

A+	0.827	0.827	0.827
A-	0.165	0.165	0.165

P+	P-
0.78	0.706
0.745	0.707
0.689	0.715
0.996	0.161
1.122	0.041

	H	Rank
P1	0.475	3
P2	0.487	2
P3	0.509	1
P4	0.139	4
P5	0.036	5

Fig 3. Illustration of TOPSIS. This illustrative example explains the functionality of TOPSIS method in a simple application in which we have 5 alternatives, P1-P5, and 3 criteria, C1-C3.

https://doi.org/10.1371/journal.pone.0184203.g003

knowing the mapping, thus high dimensionality of the feature space does not increase computational cost of classification and training task.

RF is an ensemble classifier comprised of many decision trees. The mode of class output obtained by individual trees would be the class that is output by RF[42]. The Random Decision Forests learning algorithm was developed by Leo Breiman [21] based on decision trees, which are non-parametric supervised learning approaches used for regression and classification. Using of a set of tree classifiers and randomness in the RF design led to good accuracy and stability of the resulting classifier.

RF is a classifier including a set of tree-structured classifiers $\{g(x, b_k)\ k = 1, 2,. . .\}$, where the $\{b_k\}$ are independent identically distributed random vectors and each tree puts a unit vote for the famous class at input x. The RF method (along with other ensemble learning methods) has been very popular in biomedical research, and it considers random tree building using both bagging and random variable selection [43].

Fuzzy Inference System (FIS) is a method of mapping the input space to an output space using FL. FIS attempts to formalize the reasoning procedure of human language by means of FL and building fuzzy IF-THEN rules. The procedure of fuzzy inference involves all of the sections that are explained in Membership Functions, Logical Operations, and If-Then Rules. They have become strong methods to afford various problems such as uncertainty, imprecision, and non-linearity. They are generally used for identification, classification, and regression works. Instead of employing crisp sets as in classical rules, fuzzy rules exploit fuzzy sets. Rules were initially taken from human experts through knowledge engineering procedures. However, this approach may not be possible when facing complicated tasks or when human experts are not accessible [44].

The kNN algorithm, one of the popular machine learning algorithms, is a non-parametric method used for classification and regression predictive problems. In both cases, the input vector contains the k closest training samples in the feature space. The output is dependent to value of k whether it is used for classification or regression. In kNN classification, the output is a class membership. An object is classified by a majority vote of its neighbors with the object being allocated to the class most usual between its k nearest neighbors. The best election of k depends on the data; a good k can be elected by different heuristic methods. Larger values of k decrease the effect of noise on the classification, but it creates boundaries between classes less distinct. A drawback of the kNN algorithm is its sensitivity to the local structure of the data. In kNN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors rather than voting from nearest neighbors [45].

LDA is a method used in pattern recognition, statistics, and machine learning to detect a linear combination of features that separate two or more classes of objects and is an extension of Fisher's linear discriminant; Such combination might be used as a linear classifier, or, more generally, for dimensionality reduction before later classification. LDA attempts to represent one dependent variable as a linear combination of other features and is closely relevant to analysis of variance and regression analysis [46]. LDA is closely relevant to factor analysis and principal component analysis (PCA) in that they both look for linear combinations of variables which best illustrate the data. LDA clearly efforts to model the diversity among the classes of data. PCA, on the other hand, does not take into account any diversity in class, and factor analysis creates the feature combinations according to the differences rather than similarities [46].

The Bayesian Classification is a statistical method for classification that illustrates a supervised learning strategy. Bayesian classification provides practical learning algorithms in which the former knowledge and the observed data can be combined. Bayesian Classification provides an effective perspective for evaluating and understanding many learning algorithms. It is not affected by noise in input data and calculates clear probabilities for hypothesis. The NB classifier is used when features are independent of each other within each class, but it works well in practice even when that independence assumption is not valid. NB classifier requires a small amount of training data to estimate the parameters such as mean and variance of the variables necessary for classification [47].

Performance measures

K-fold cross-validation test, independent dataset test, sub-sampling test and jackknife cross-validation test are four widely used classes of schemes in statistical classify for examining the performance of a prediction model [48–53]. The jackknife test has been widely used in Bioinformatics [54–68], because it can achieve unique outcome [33, 69]. However, it is time-consuming. For saving the computational time, in this study, ten-fold cross-validation was used to investigate the performance of the prediction model. In k-fold cross-validation, the data is divided into k subset, each time, one of the k subsets and k-1 subsets are used as test and train data, respectively. Then the mean error across all k experiment is calculated. Since the utilized dataset is unbalanced in terms of number of samples in two groups of early and advanced stages, the Area under Curve (AUC) and Matthews Correlation Coefficient (MCC) were used.

The MCC, introduced by Brian W. Matthews [70], is used for measuring the quality of binary classification. The MCC is a number between -1 and +1. Values of 1 and 0 demonstrate a perfect and random prediction, respectively. In addition, -1 represents total disagreement between the predicted and actual values. It can be computed from the confusion matrix as:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{11}$$

where TP is the number of true positives (early stage), TN is the number of true negatives (advanced stage), FP is the number of false positives, and FN is the number of false negatives. Eq 11 can be represented in another form like Eq. 11 in [40], which were derived by Xu et al. [29] and Lin et al. [40] based on the symbols introduced by Chou in studying signal peptides and those used in many recent studies [26–32]. The set of metrics is valid only for the single-label systems. For the multi-label systems which has become more frequent in systems biology [71] and systems medicine [37, 72–74], a completely different set of metrics as defined in [75] is needed.

Moreover, the AUC is defined as the area under the ROC curve, which illustrates the performance of a binary classifier system as its discrimination threshold is varied. An AUC of 1, 0.5, and under 0.5 indicates a perfect, random, and bad classifier, respectively.

Results

As can be seen from the data in Table 1, the KIRC cancer with 453 samples, contains the most samples in the whole dataset. The UCEC Cancer is the second-most with 404 samples. According to the pathologic stage, data are unbalanced and the READ and OV data has the less and the most unbalanced level, respectively.

To present the performance of the proposed FRMT method, we have provided 7 tables; one table for each cancer data (S1 Table), which demonstrated the effect of applying different feature selection technique on various classifier architectures. In this regard, MCC and AUC are used as evaluation measures in 30 repetitions of 10-fold cross-validation procedure. For a fair performance evaluation, we should consider different constraints that affect the classification performance such as: train dataset, classifier model, and the number of selected features. In this regard, we should evaluate different possible combinations, which contains 49 states due to the seven classifiers that applied on seven datasets. Then, we select subset of features with

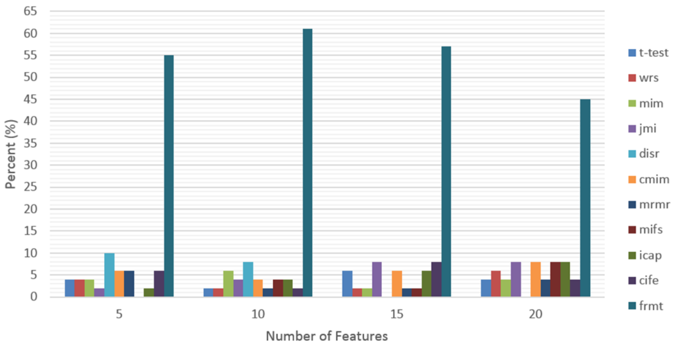


Fig 4. Effect of feature subset size on performance. The winning frequency is calculated for different feature selection methods for various sizes of feature subsets.

https://doi.org/10.1371/journal.pone.0184203.g004

different sizes (5, 10, 15 and 20) obtained by each feature selection method, considering which method reaches the highest accuracy in each of 49 state. Fig 4 shows the percentage of states that each feature selection method reached the best performance (winning frequency) for various numbers of features. As it is shown in Fig 4, the proposed method has reached the best result for all sizes of feature subsets in comparison with other methods, and the peak of result obtained by using 10 features.

After this point the same number of features (top 10 proteins) has been selected as the input of all classifiers in all experiments. The best results in the tables are highlighted by shading. The frequency of selection of each feature selection method as the best, or winning frequency regarding the classification performance, is depicted in Fig 5. For each cancer, each classifier model obtained the best answer with only one of the eleven feature selection methods.

Fig 6 illustrates the winning frequency of all feature selection methods in the whole dataset. The method with larger segment on the pie chart demonstrates the better approach.

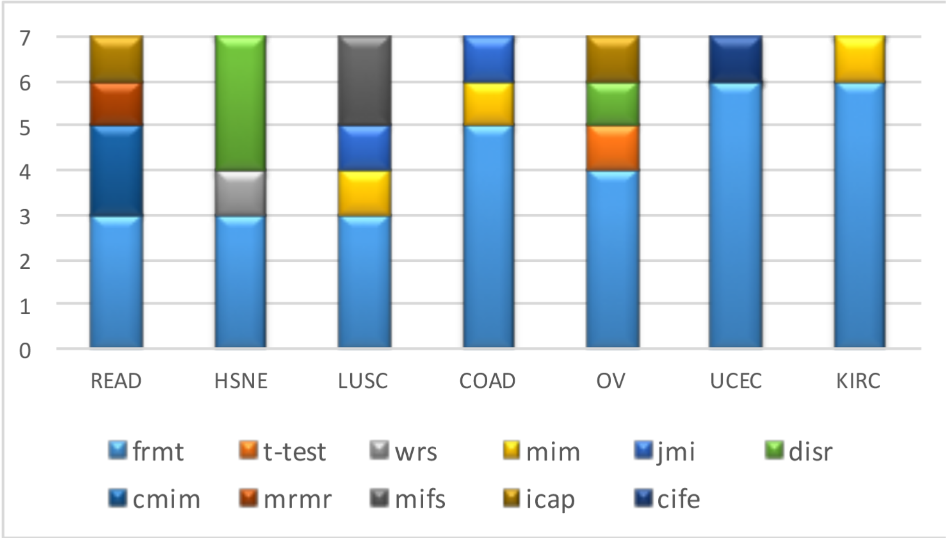


Fig 5. Comparison of FFS methods in different datasets. The winning frequency is calculated for different feature selection methods in each dataset, regarding the classification performance.

https://doi.org/10.1371/journal.pone.0184203.g005

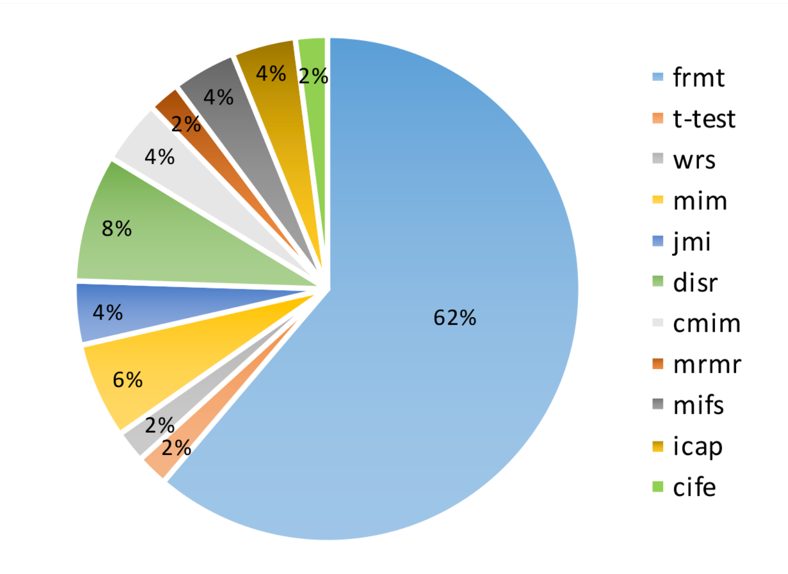


Fig 6. Final comparison of different FFS methods. The winning frequency of all feature selection methods is illustrated in a pie chart for all datasets.

https://doi.org/10.1371/journal.pone.0184203.g006

The results presented in S1 Table are summarized in Table 3. The left part of Table 3 demonstrates a comparative analysis of the FRMT method performance by applying different classification models in seven datasets. In the right part of Table 3, the best results of every feature selection method in combination with a classifier that led to the best performance for prediction of cancer stage are shown.

After applying the FRMT method in different datasets, top ranked proteins were extracted and the name of first 10 informative ones were reported in S1 Table.

Discussion

In this study, a new approach called FRMT method was proposed to select protein biomarkers. 10 FFS methods were integrated to extract the best stage prediction cancer biomarkers. Finding the best proteins via a multi-criteria decision analysis method, the FRMT method demonstrates a proficient method for ranking proteins using protein expression profile data without concerning about the selection of suitable FFS method for a specific problem.

Table 3. Comparison of the FRMT method with other methods for whole cancers.

	FRMT method			Other methods			
Cancer	MCC	AUC	Classifier	MCC	AUC	Criteria	Classifier
READ	**0.23±0.0025**	**61.59±0.61**	LDA	**0.23±0.0021**	59.09±0.31	*cmim*	SVM
HSNE	**0.37±0.0002**	**69.55±0.07**	NB	0.27±0.0018	60.59±0.31	*disr*	KNN
LUSC	**0.27±0.002**	**59.02±0.28**	KNN	0.16±0.0068	57.93±1.69	*mim*	DT
COAD	**0.18±0.0015**	**58.55±0.31**	SVM	0.15±0.0013	56.97±0.3	*mim*	SVM
OV	**0.27±0.0007**	**61.57±0.11**	NB	0.17±0.0013	54.58±0.12	*cife*	NB
UCEC	**0.26±0.0011**	**56.54±0.08**	NB	0.16±0.0005	55.96±0.08	*disr*	NB
KIRC	**0.34±0.0001**	**66.54±0.03**	NB	0.28±0.0005	63.51±0.13	*wrs*	RF

https://doi.org/10.1371/journal.pone.0184203.t003

The performance of six well-known classifiers was evaluated and reported using 10 top ranked proteins selected by the FRMT and other FFS approaches. The results indicate that the FRMT method is more advantageous than other FFS methods in terms of robustness in classification performance; By measuring the number of times that a method obtained the best results, we observed that the best frequency has been achieved by the FRMT method with 6 out of 7 times in UCEC and KIRC cancer (Fig 5). Furthermore, in the READ and LUSC cancer dataset, the maximum frequency of 3 out of 7 has been reached by the FRMT method. In the HSNE cancer, the frequency for *disr* and FRMT methods are equal to 3. It should be noted that some methods of feature selection were never chosen as the best model.

Looking at the pie chart in Fig 6, in 62 percent of experiments the FRMT method achieved the best classification performance in the whole proteomic dataset. Afterward, the *disr* and *mim* methods reached to success rate of 8 and 6 percent, respectively.

As it is reported in Table 3, the best performance in prediction of cancer stage evaluated by using AUC and MCC has occurred in HSNE dataset. The AUC of 69.55 with SE (Standard Error) of 0.07 and MCC of 0.37 with SE of 0.0002 are the best results among all dataset achieved by the FRMT method as feature selection and NB method as classifier. It should be noted that in HSNE dataset, the *disr* method reached the second-best place by AUC of 60.59 with SE of 0.31 and MCC of 0.27 with SE of 0.0018.

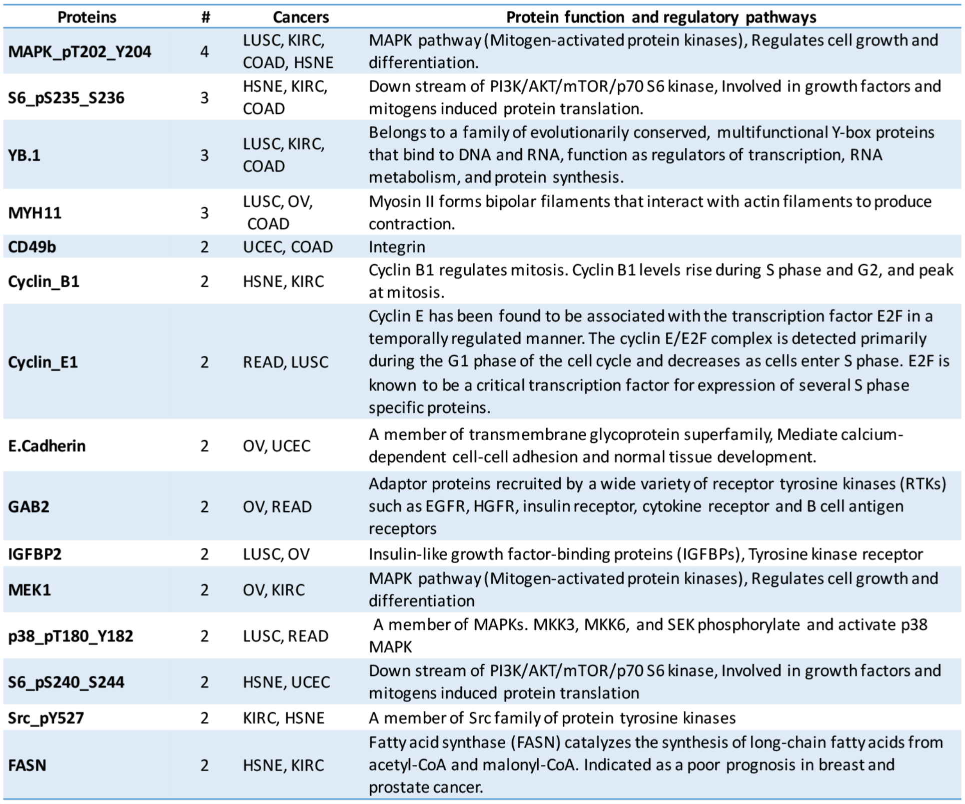
Comparison of the FRMT method with other methods in Table 3 are suggestive of the *wrs* method achieving the best results among other FFS methods; AUC of 63.51 with SE of 0.13 and MCC of 0.28 with SE of 0.0005 have been achieved by using RF as classifier in KIRC dataset. However, in the KIRC dataset, the FRMT method already obtained the best result with NB classifier, which are AUC of 66.54 with SE of 0.03 and MCC of 0.34 with SE of 0.0001.

As it is seen from the data in Table 3, the NB classifier was achieved the best results in the majority of experiments evaluated various feature selection methods. NB achieved the best performance in 4 out of 7 datasets using the proposed method, and in two datasets by applying *cife* and *disr* methods. Notably, the SVM classifier obtained the second place.

Top-ranked protein selected by FRMT method from each dataset, showed significant overlap with recently discovered biomarkers that were associated with cancer development. According to Fig 7, MAPK_pT202_Y204 is the most frequently selected protein from 4 datasets among top 10 ones by FRMT. The striking point about the MAPK_pT202_Y204 is its significant role in MAPK pathway (Mitogen-activated protein kinases) and regulation of cell growth and differentiation [34].

In addition, S6_pS235_S236 which involved in growth factors and mitogens induced protein translation [34], is the second frequently selected protein selected from 3 datasets among top 10 ones by FRMT.

Gab2 that is selected by FRMT as the most informative protein in the READ dataset is recently introduced as an overexpressed protein in several cancer types [76–78]. Moreover, several researchers have reported that overexpression of Gab2 stimulates cell proliferation, cell transformation, and tumor progression; Ding et al. [79] showed Gab2 overexpression in clinical colorectal cancer (CRC) specimens. Moreover, Gab2 is selected by FRMT as the second discriminative protein in OV dataset, and this is in concordance with recent studies that reported Gab2 amplification and overexpression in a subset of primary high-grade serous ovarian cancers and cell lines [78]. Furthermore, the expression level of IRS1, which is selected by FRMT as the second discriminative protein in READ dataset, was utilized by Hanyuda et al. as a predictive marker for classification of patients according to their survival benefit gained by the exercise [80].



Proteins	#	Cancers	Protein function and regulatory pathways
MAPK_pT202_Y204	4	LUSC, KIRC, COAD, HSNE	MAPK pathway (Mitogen-activated protein kinases), Regulates cell growth and differentiation.
S6_pS235_S236	3	HSNE, KIRC, COAD	Down stream of PI3K/AKT/mTOR/p70 S6 kinase, Involved in growth factors and mitogens induced protein translation.
YB.1	3	LUSC, KIRC, COAD	Belongs to a family of evolutionarily conserved, multifunctional Y-box proteins that bind to DNA and RNA, function as regulators of transcription, RNA metabolism, and protein synthesis.
MYH11	3	LUSC, OV, COAD	Myosin II forms bipolar filaments that interact with actin filaments to produce contraction.
CD49b	2	UCEC, COAD	Integrin
Cyclin_B1	2	HSNE, KIRC	Cyclin B1 regulates mitosis. Cyclin B1 levels rise during S phase and G2, and peak at mitosis.
Cyclin_E1	2	READ, LUSC	Cyclin E has been found to be associated with the transcription factor E2F in a temporally regulated manner. The cyclin E/E2F complex is detected primarily during the G1 phase of the cell cycle and decreases as cells enter S phase. E2F is known to be a critical transcription factor for expression of several S phase specific proteins.
E.Cadherin	2	OV, UCEC	A member of transmembrane glycoprotein superfamily, Mediate calcium-dependent cell-cell adhesion and normal tissue development.
GAB2	2	OV, READ	Adaptor proteins recruited by a wide variety of receptor tyrosine kinases (RTKs) such as EGFR, HGFR, insulin receptor, cytokine receptor and B cell antigen receptors
IGFBP2	2	LUSC, OV	Insulin-like growth factor-binding proteins (IGFBPs), Tyrosine kinase receptor
MEK1	2	OV, KIRC	MAPK pathway (Mitogen-activated protein kinases), Regulates cell growth and differentiation
p38_pT180_Y182	2	LUSC, READ	A member of MAPKs. MKK3, MKK6, and SEK phosphorylate and activate p38 MAPK
S6_pS240_S244	2	HSNE, UCEC	Down stream of PI3K/AKT/mTOR/p70 S6 kinase, Involved in growth factors and mitogens induced protein translation
Src_pY527	2	KIRC, HSNE	A member of Src family of protein tyrosine kinases
FASN	2	HSNE, KIRC	Fatty acid synthase (FASN) catalyzes the synthesis of long-chain fatty acids from acetyl-CoA and malonyl-CoA. Indicated as a poor prognosis in breast and prostate cancer.

Fig 7. Detail description of top ranked proteins. The name, frequency of selection, related cancer, function and regulatory pathways of informative proteins are reported, which are appeared more than one time in whole cancers among the 10 top ranking of selected proteins by FRMT method.

https://doi.org/10.1371/journal.pone.0184203.g007

About S6 phosphorylation(S6_pS240_S244), which is selected by FRMT as the most discriminative protein in the HSNE dataset, previous studies have revealed its high occurrence in HNSCC specimens and demonstrated its correlation on clinical outcomes [81].

Bcl-2 protein is chosen by FRMT as an important marker in the COAD dataset; This finding broadly supports the work of Poincloux et al., linking loss of Bcl-2 protein expression with increase in relapse of stage II colon cancer, and it could be a potential histo-prognostic marker in therapy decision making [82].

Many studies [53, 60–62, 83–85] have demonstrated that high dimension data will bring about information redundancy or noise that results in bad prediction accuracy, over-fitting that results in low generalization ability of prediction model, and dimension disaster which in turn is a handicap for the computation. Thus, a novel two-step feature selection technique was applied to optimize features.

As demonstrated in a series of recent publications (see, e.g., [26–32, 86, 87]) in evaluating new prediction/classification methods, user-friendly and publicly accessible web-servers will significantly enhance their impacts [88, 89], we shall try to provide a web-server in our future work for online application of the method presented in this paper. Moreover, for extending our experiment, we shall consider combining different feature selector as in [90].

Conclusion

Various FFS methods may lead to diverse biomarkers with different discriminative power in different datasets. However, the proposed FRMT method can help researchers to select more stable biomarkers from protein expression profiles by integrating various FFS methods. The proposed method has the advantage of stability and classification performance compared with other approaches. However, it suffers from the computational complexity problem comparing to FFS methods. On the other hand, the FRMT method in comparison to the wrapper feature selection approaches, has lower computational complexity and produce more general results without overfitting.

Supporting information

S1 Table. Name of proteins and their classification performance in all datasets. The results obtained for 10 formative biomarkers selected by FRMT and reported in different sheets for each dataset.
(XLSX)

Acknowledgments

This research was financially supported by the Isfahan University of Medical Sciences, Isfahan, Iran (Grant no.395775). The authors would like to acknowledge the products and services provided by Isfahan University of Medical Sciences. The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/; for this reason, the authors would like to greatly acknowledge the TCGA Research Network. Also, we wish to acknowledge MS Solutions LLC members for supplying us with editing, proofreading, and high-quality research publication support services.

Author Contributions

Conceptualization: Ehsan Saghapour, Saeed Kermani, Mohammadreza Sehhati.

Formal analysis: Ehsan Saghapour.

Funding acquisition: Saeed Kermani.

Methodology: Ehsan Saghapour.

Project administration: Saeed Kermani.

Software: Ehsan Saghapour.

Supervision: Saeed Kermani, Mohammadreza Sehhati.

Validation: Ehsan Saghapour.

Visualization: Ehsan Saghapour.

Writing – original draft: Ehsan Saghapour, Mohammadreza Sehhati.

Writing – review & editing: Ehsan Saghapour, Saeed Kermani, Mohammadreza Sehhati.

References

1. Lu J-W, Shen C, Tzeng T-Y. Epigenetics of cancer: the role of histone methyltransferase, SETDB1, in cancer metastasis. AME PUBL CO ROOM 604 6-F HOLLYWOOD CENTER, 77–91, QUEENS ROAD, SHEUNG WAN, HONG KONG 00000, PEOPLES R CHINA; 2016.

2. Azodi MZ, Ardestani H, Dolat E, Mousavi M, Fayazfar S, Shadloo A. Breast cancer: Genetics, risk factors, molecular pathology and treatment. Journal of Paramedical Sciences. 2012; 4(1).

3. Khatib H, Rezaei-Tavirani M, Keshel SH, Azodi MZ, Omidi R, Biglarian M, et al. Flow cytometry analysis of Rosa Damascena effects on gastric cancer cell line (MKN45). Iranian Journal of Cancer Prevention. 2013; 6:30–6.

4. Rezaie-Tavirani M, Fayazfar S, Heydari-Keshel S, Rezaee MB, Zamanian-Azodi M, Rezaei-Tavirani M, et al. Effect of essential oil of Rosa Damascena on human colon cancer cell line SW742. Gastroenterology and Hepatology from bed to bench. 2013; 6(1).

5. Zali H, Rezaei-Tavirani M, Azodi M. Gastric cancer: prevention, risk factors and treatment. Gastroenterology and Hepatology from bed to bench. 2011; 4(4).

6. Honda K, Ono M, Shitashige M, Masuda M, Kamita M, Miura N, et al. Proteomic approaches to the discovery of cancer biomarkers for early detection and personalized medicine. Japanese journal of clinical oncology. 2012:hys200.

7. Saghapour E, Sehhati M. Prediction of metastasis in advanced colorectal carcinomas using CGH data. Journal of Theoretical Biology. 2017.

8. Mazumder A, Palma AJF, Wang Y. Validation and integration of gene-expression signatures in cancer. Expert review of molecular diagnostics. 2008; 8(2):125–8. https://doi.org/10.1586/14737159.8.2.125 PMID: 18366298

9. Sehhati M, Mehridehnavi A, Rabbani H, Pourhossein M. Stable Gene Signature Selection for Prediction of Breast Cancer Recurrence Using Joint Mutual Information. IEEE/ACM transactions on computational biology and bioinformatics. 2015; 12(6):1440–8. https://doi.org/10.1109/TCBB.2015.2407407 PMID: 26671813

10. Zhang P-W, Chen L, Huang T, Zhang N, Kong X-Y, Cai Y-D. Classifying ten types of major cancers based on reverse phase protein array profiles. PloS one. 2015; 10(3):e0123147. https://doi.org/10.1371/journal.pone.0123147 PMID: 25822500

11. Sonntag J, Bender C, Soons Z, von der Heyde S, König R, Wiemann S, et al. Reverse phase protein array based tumor profiling identifies a biomarker signature for risk classification of hormone receptor-positive breast cancer. Advances in Integrative Medicine. 2014; 2:52–9.

12. Kaddi C, Wang M. Models for Predicting Stage in Head and Neck Squamous Cell Carcinoma using Proteomic and Transcriptomic Data. 2015.

13. Stafford P, Cichacz Z, Woodbury NW, Johnston SA. Immunosignature system for diagnosis of cancer. Proceedings of the National Academy of Sciences. 2014; 111(30):E3072–E80.

14. Hanash S, Taguchi A. The grand challenge to decipher the cancer proteome. Nature reviews cancer. 2010; 10(9):652–60. https://doi.org/10.1038/nrc2918 PMID: 20733593

15. Nguyen T, Nahavandi S. Modified AHP for Gene Selection and Cancer Classification Using Type-2 Fuzzy Logic. IEEE Transactions on Fuzzy Systems. 2016; 24(2):273–87.

16. Nguyen T, Khosravi A, Creighton D, Nahavandi S. Hierarchical Gene Selection and Genetic Fuzzy System for Cancer Microarray Data Classification. PloS one. 2015; 10(3):e0120364. https://doi.org/10.1371/journal.pone.0120364 PMID: 25823003

17. Karnik NN, Mendel JM, Liang Q. Type-2 fuzzy logic systems. IEEE transactions on Fuzzy Systems. 1999; 7(6):643–58.

18. Eddy SR. Hidden markov models. Current opinion in structural biology. 1996; 6(3):361–5. PMID: 8804822

19. Ritter G, Woodruff H, Lowry S, Isenhour T. An algorithm for a selective nearest neighbor decision rule. IEEE Transactions on Information Theory. 1975; 21(6):665–9.

20. Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications. 1998; 13(4):18–28.

21. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32.

22. Quinlan JR. Simplifying decision trees. International journal of man-machine studies. 1987; 27(3): 221–34.

23. Scholkopft B, Mullert K-R. Fisher discriminant analysis with kernels. Neural networks for signal processing IX. 1999; 1(1):1.

24. Duda RO, Hart PE, Stork DG. Pattern classification: John Wiley & Sons; 2012.

25. Klir G, Yuan B. Fuzzy sets and fuzzy logic: Prentice hall New Jersey; 1995.

26. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of theoretical biology. 2016; 394:223–30. https://doi.org/10.1016/j.jtbi.2016.01.020 PMID: 26807806

27. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget. 2016; 7(23): 34558. https://doi.org/10.18632/oncotarget.9148 PMID: 27153555

28. Qiu W-R, Jiang S-Y, Xu Z-C, Xiao X, Chou K-C. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget. 2017.

29. Xu Y, Ding J, Wu L-Y, Chou K-C. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One. 2013; 8(2):e55844. https://doi.org/10.1371/journal.pone.0055844 PMID: 23409062

30. Xu Y, Shao X-J, Wu L-Y, Deng N-Y, Chou K-C. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ. 2013; 1:e171. https://doi.org/10.7717/peerj.171 PMID: 24109555

31. Xu Y, Wen X, Shao X-J, Deng N-Y, Chou K-C. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. International journal of molecular sciences. 2014; 15(5):7594–610. https://doi.org/10.3390/ijms15057594 PMID: 24857907

32. Xu Y, Wen X, Wen L-S, Wu L-Y, Deng N-Y, Chou K-C. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PloS one. 2014; 9(8):e105018. https://doi.org/10.1371/journal.pone.0105018 PMID: 25121969

33. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology. 2011; 273(1):236–47. https://doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

34. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for cancer functional proteomics data. Nature methods. 2013; 10(11):1046–7.

35. Brown G, Pocock A, Zhao M-J, Luján M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. Journal of Machine Learning Research. 2012; 13(Jan):27–66.

36. Cai Y-D, Zhou G-P, Chou K-C. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. Journal of theoretical biology. 2005; 234(1):145–9. https://doi.org/10.1016/j.jtbi.2004.11.017 PMID: 15721043

37. Cheng X, Zhao S, Xiao X, Chou K. iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget. 2017.

38. Xiao X, Wang P, Chou K-C. Predicting the quaternary structure attribute of a protein by hybridizing functional domain composition and pseudo amino acid composition. Journal of Applied Crystallography. 2009; 42(2):169–73.

39. Xiao X, Wang P, Chou K-C. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Molecular Biosystems. 2011; 7(3):911–9. https://doi.org/10.1039/c0mb00170h PMID: 21180772

40. Lin H, Deng E-Z, Ding H, Chen W, Chou K-C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic acids research. 2014; 42(21):12961–72. https://doi.org/10.1093/nar/gku1019 PMID: 25361964

41. Tzeng G-H, Huang J-J. Multiple attribute decision making: methods and applications: CRC press; 2011.

42. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3):273–97.

43. Ho TK, editor Random decision forests. Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on; 1995: IEEE.

44. Pelletier FJ. Review of Metamathematics of fuzzy logics in The Bulletin of Symbolic Logic, Vol. 6, No. 3, (Sep. 2000), 342–346. JSTOR.

45. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992; 46(3):175–85.

46. McLachlan G. Discriminant analysis and statistical pattern recognition: John Wiley & Sons; 2004.

47. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.

48. Ding H, Luo L, Lin H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein and peptide letters. 2009; 16(4):351–5. PMID: 19356130

49. Ding H, Li D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino acids. 2015; 47(2):329–33. https://doi.org/10.1007/s00726-014-1862-4 PMID: 25385313

50. Lin H, Ding H, Guo F-B, Zhang A-Y, Huang J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein and peptide letters. 2008; 15(7):739–44. PMID: 18782071

51. Lin H, Ding C, Song Q, Yang P, Ding H, Deng K-J, et al. The prediction of protein structural class using averaged chemical shifts. Journal of Biomolecular Structure and Dynamics. 2012; 29(6):1147–53.

52. Lin H, Li Q-Z. Eukaryotic and prokaryotic promoter prediction using hybrid approach. Theory in Biosciences. 2011; 130(2):91–100. https://doi.org/10.1007/s12064-010-0114-8 PMID: 21046474

53. Zhao Y-W, Lai H-Y, Tang H, Chen W, Lin H. Prediction of phosphothreonine sites in human proteins by fusing different features. Scientific reports. 2016; 6.

54. Nanni L, Brahnam S, Lumini A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. Journal of theoretical biology. 2014; 360:109–16. https://doi.org/10.1016/j.jtbi.2014.07.003 PMID: 25026218

55. Behbahani M, Mohabatkar H, Nosrati M. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. Journal of theoretical biology. 2016; 411:1–5. https://doi.org/10.1016/j.jtbi.2016.09.001 PMID: 27615149

56. Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Scientific Reports. 2017; 7.

57. Tripathi P, Pandey PN. A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. Journal of Theoretical Biology. 2017; 424:49–54. https://doi.org/10.1016/j.jtbi.2017.04.027 PMID: 28476562

58. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001; 43(3):246–55.

59. Lin H, Ding H, Guo F-B, Huang J. Prediction of subcellular location of mycobacterial protein using feature selection techniques. Molecular diversity. 2010; 14(4):667–71. https://doi.org/10.1007/s11030-009-9205-1 PMID: 19908156

60. Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. Journal of microbiological methods. 2011; 84(1):67–70. https://doi.org/10.1016/j.mimet.2010.10.013 PMID: 21044646

61. Yuan L-F, Ding C, Guo S-H, Ding H, Chen W, Lin H. Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. Toxicology in Vitro. 2013; 27(2):852–6. https://doi.org/10.1016/j.tiv.2012.12.024 PMID: 23280100

62. Ding H, Feng P-M, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Molecular BioSystems. 2014; 10(8):2229–35. https://doi.org/10.1039/c4mb00316k PMID: 24931825

63. Ding H, Lin H, Chen W, Li Z-Q, Guo F-B, Huang J, et al. Prediction of protein structural classes based on feature selection technique. Interdisciplinary sciences, computational life sciences. 2014; 6(3):235. https://doi.org/10.1007/s12539-013-0205-6 PMID: 25205501

64. Ding H, Liang Z-Y, Guo F-B, Huang J, Chen W, Lin H. Predicting bacteriophage proteins located in host cell with feature selection technique. Computers in biology and medicine. 2016; 71:156–61. https://doi.org/10.1016/j.compbiomed.2016.02.012 PMID: 26945463

65. Ding H, Yang W, Tang H, Feng P-M, Huang J, Chen W, et al. PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. Virologica Sinica. 2016; 31(4):350. https://doi.org/10.1007/s12250-016-3740-6 PMID: 27151186

66. Tang H, Zou P, Zhang C, Chen R, Chen W, Lin H. Identification of apolipoprotein using feature selection technique. Scientific reports. 2016; 6.

67. Tang H, Su Z-D, Wei H-H, Chen W, Lin H. Prediction of cell-penetrating peptides with feature selection techniques. Biochemical and biophysical research communications. 2016; 477(1):150–4. https://doi.org/10.1016/j.bbrc.2016.06.035 PMID: 27291150

68. Lai H-Y, Chen X-X, Chen W, Tang H, Lin H. Sequence-based predictive modeling to identify cancerlectins. Oncotarget. 2017; 8(17):28169. https://doi.org/10.18632/oncotarget.15963 PMID: 28423655

69. Chou K-C, Zhang C-T. Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology. 1995; 30(4):275–349. https://doi.org/10.3109/10409239509083488 PMID: 7587280

70. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1975; 405(2):442–51.

71. Chou K-C, Wu Z-C, Xiao X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Molecular Biosystems. 2012; 8(2):629–41. https://doi.org/10.1039/c1mb05420a PMID: 22134333

72. Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Analytical biochemistry. 2013; 436(2):168–77. https://doi.org/10.1016/j.ab.2013.01.019 PMID: 23395824

73. Qiu W-R, Sun B-Q, Xiao X, Xu Z-C, Chou K-C. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics. 2016; 32(20):3116–23. https://doi.org/10.1093/bioinformatics/btw380 PMID: 27334473

74. Cheng X, Zhao S-G, Xiao X, Chou K-C. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics. 2016; 33(3):341–6.

75. Chou K-C. Some remarks on predicting multi-label attributes in molecular biosystems. Molecular Biosystems. 2013; 9(6):1092–100. https://doi.org/10.1039/c3mb25555g PMID: 23536215

76. Ding C, Luo J, Li L, Li S, Yang L, Pan H, et al. Gab2 facilitates epithelial-to-mesenchymal transition via the MEK/ERK/MMP signaling in colorectal cancer. Journal of Experimental & Clinical Cancer Research. 2016; 35(1):1.

77. Fleuren ED, O'Toole S, Millar EK, McNeil C, Lopez-Knowles E, Boulghourjian A, et al. Overexpression of the oncogenic signal transducer Gab2 occurs early in breast cancer development. International journal of cancer. 2010; 127(6):1486–92. https://doi.org/10.1002/ijc.25172 PMID: 20087860

78. Duckworth C, Zhang L, Carroll S, Ethier S, Cheung H. Overexpression of GAB2 in ovarian cancer cells promotes tumor growth and angiogenesis by upregulating chemokine expression. Oncogene. 2015.

79. Ding C, Luo J, Yu W, Gao S, Yang L, Chen C, et al. Gab2 is a novel prognostic factor for colorectal cancer patients. Int J Clin Exp Pathol. 2015; 8(3):2779–86. PMID: 26045784

80. Hanyuda A, Kim SA, Martinez-Fernandez A, Qian ZR, Yamauchi M, Nishihara R, et al. Survival Benefit of Exercise Differs by Tumor IRS1 Expression Status in Colorectal Cancer. Annals of surgical oncology. 2016; 23(3):908–17. https://doi.org/10.1245/s10434-015-4967-4 PMID: 26577117

81. García-Carracedo D, Angeles Villaronga M, Álvarez-Teijeiro S, Hermida-Prado F, Santamaría I, Allonca E, et al. Impact of PI3K/AKT/mTOR pathway activation on the prognosis of patients with head and neck squamous cell carcinomas. Oncotarget. 2016; 7(20):29780–93. https://doi.org/10.18632/oncotarget.8957 PMID: 27119232

82. Poincloux L, Durando X, Seitz JF, Thivat E, Bardou V-J, Giovannini M-H, et al. Loss of Bcl-2 expression in colon cancer: a prognostic factor for recurrence in stage II colon cancer. Surgical oncology. 2009; 18(4):357–65. https://doi.org/10.1016/j.suronc.2008.09.003 PMID: 19027288

83. Wei L, Xing P, Shi G, Ji Z-L, Zou Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2017.

84. Yang H, Tang H, Chen X-X, Zhang C-J, Zhu P-P, Ding H, et al. Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. BioMed research international. 2016; 2016.

85. Chen X-X, Tang H, Li W-C, Wu H, Chen W, Ding H, et al. Identification of bacterial cell wall lyases via pseudo amino acid composition. BioMed research international. 2016; 2016.

86. Liu L, Xu Y, Chou K. iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Medicinal chemistry (Shariqah (United Arab Emirates)). 2017.

87. Xu Y, Wang Z, Li C, Chou K. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Medicinal chemistry (Shariqah (United Arab Emirates)). 2017.

88. Chou K-C. Impacts of bioinformatics to medicinal chemistry. Medicinal chemistry. 2015; 11(3):218–34. PMID: 25548930

89. Liu B, Wu H, Zhang D, Wang X, Chou K-C. Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. Oncotarget. 2017; 8(8):13338. https://doi.org/10.18632/oncotarget.14524 PMID: 28076851

90. Nanni L, Salvatore C, Cerasa A, Castiglioni I, Initiative AsDN. Combining multiple approaches for the early diagnosis of Alzheimer's Disease. Pattern Recognition Letters. 2016; 84:259–66.