Computational and Structural Biotechnology Journal 18 (2020) 2012–2025

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Review

Computational advances of tumor marker selection and sample classification in cancer proteomics

Jing Tang [a,b], Yunxia Wang [b], Yongchao Luo [b], Jianbo Fu [b], Yang Zhang [b,c], Yi Li [b], Ziyu Xiao [b], Yan Lou [d], Yunqing Qiu [d,*], Feng Zhu [a,b,*]

[a] Department of Bioinformatics, Chongqing Medical University, Chongqing 400016, China
[b] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
[c] School of Pharmaceutical Sciences and Innovative Drug Research Centre, Chongqing University, Chongqing 401331, China
[d] Zhejiang Provincial Key Laboratory for Drug Clinical Research and Evaluation, The First Affiliated Hospital, Zhejiang University, Hangzhou 310000, China

ARTICLE INFO

Article history:
Received 7 April 2020
Received in revised form 6 July 2020
Accepted 8 July 2020
Available online 17 July 2020

Keywords:
Cancer proteomics
Tumor marker selection
Sample classification
Computational methods

ABSTRACT

Cancer proteomics has become a powerful technique for characterizing the protein markers driving transformation of malignancy, tracing proteome variation triggered by therapeutics, and discovering the novel targets and drugs for the treatment of oncologic diseases. To facilitate cancer diagnosis/prognosis and accelerate drug target discovery, a variety of methods for tumor marker identification and sample classification have been developed and successfully applied to cancer proteomic studies. This review article describes the most recent advances in those various approaches together with their current applications in cancer-related studies. Firstly, a number of popular feature selection methods are overviewed with objective evaluation on their advantages and disadvantages. Secondly, these methods are grouped into three major classes based on their underlying algorithms. Finally, a variety of sample separation algorithms are discussed. This review provides a comprehensive overview of the advances on tumor maker identification and patients/samples/tissues separations, which could be guidance to the researches in cancer proteomics.

Contents

Abbreviations: ANN, Artificial Neural Network; ANOVA, Analysis of Variance; CFS, Correlation-based Feature Selection; DAPC, Discriminant Analysis of Principal Component; DT, Decision Trees; EDA, Estimation of Distribution Algorithm; FC, Fold Change; GA, Genetic Algorithms; GR, Gain Ratio; HC, Hill Climbing; HCA, Hierarchical Cluster Analysis; IG, Information Gain; LDA, Linear Discriminant Analysis; LIMMA, Linear Models for Microarray Data; MBF, Markov Blanket Filter; MWW, Mann–Whitney–Wilcoxon test; OPLS-DA, Orthogonal Partial Least Squares Discriminant Analysis; PCA, Principal Component Analysis; PLS-DA, Partial Least Square Discriminant Analysis; RF-RFE, Random Forest with Recursive Feature Elimination; RF, Random Forest; SA, Simulated Annealing; SAM, Significance Analysis of Microarrays; SBE, Sequential Backward Elimination; SFS, and Sequential Forward Selection; SOM, Self-organizing Map; SU, Symmetrical Uncertainty; SVM-RFE, Support Vector Machine with Recursive Feature Elimination; SVM, Support Vector Machine; sPLSDA, Sparse Partial Least Squares Discriminant Analysis; t-SNE, Student t Distribution; χ^2, Chi-square.

* Corresponding authors at: College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China (F. Zhu).
E-mail addresses: qiuyq@zju.edu.cn (Y. Qiu), zhufeng@zju.edu.cn (F. Zhu).

https://doi.org/10.1016/j.csbj.2020.07.009

1. Introduction

In 2015, cancer caused over eight million deaths, making it the second leading cause of death in the world, nearly one-sixth deaths caused by cancer according to the report of the World Health Organization [1,2]. The impact of cancer on economy is significant and growing [3–5], and the cancer is common in late manifestations and inaccessible diagnosis and treatment [6]. If detected early, cancer can respond to effective treatments, leading to lower morbidity, less treatment spending and larger chances of survival [7]. Thus, an effective early diagnosis and treatment can improve the cure rate of many cancers and alleviate the burden of cancer patients [8].

It is important to develop more effective early diagnosis strategies for cancer, and biomarkers are urgently needed to diagnose various types of cancers, assess the severity of diseases, and discover the corresponding therapeutics [9–15]. A biomarker can be a protein, a polypeptide or a metabolite whose levels change with the stage of cancer, as well as the messenger RNA or other kinds of nucleic acids [16–19]. It is well known that there is much less understanding of the pathogenesis of cancer at the proteomic level than gene mutations level [20,21]. Since proteome is a functional translation of genome and a rich source of biomarkers, huge amounts of time and money are needed for proteomics to develop biomarker [22]. Gene is merely a "formulation" of the cell, and the protein encoded by the gene is ultimately a functional participant in normal and cancer physiology [23]. Thus, while the genetic background contributes in part to the susceptibility and development of cancer [24], cancer can now be considered as a proteome disease and has more links to the post-transcriptional steps [25,26]. In terms of patient prognosis, there is an urgent need for protein markers distinguishing cancer patients from normal individuals. In addition, biomarkers used in cancer surveillance may affect the development of new therapy [27]. Protein biomarkers obtained from biological samples can not only provide starting point for finding links between disease and biological pathway, but also play an important role in advancing cancer medical research through the early diagnosis of cancer and prognosis of treatment interventions [28]. An ideal biomarker should be with the following characteristics: (1) good consistency and (2) high reproducibility on same phenotyping biological samples, and (3) good classification performance at distinguishing cases from controls across studies [29–31].

Recently, a well-rounded analysis of human genome, transcriptome, proteome and metabolome has made significant advance, making a significant contribution to the discovery of tumor biomarkers [32]. Proteomics can provide the qualitative and quantitative information on proteins based on a multitude of complex

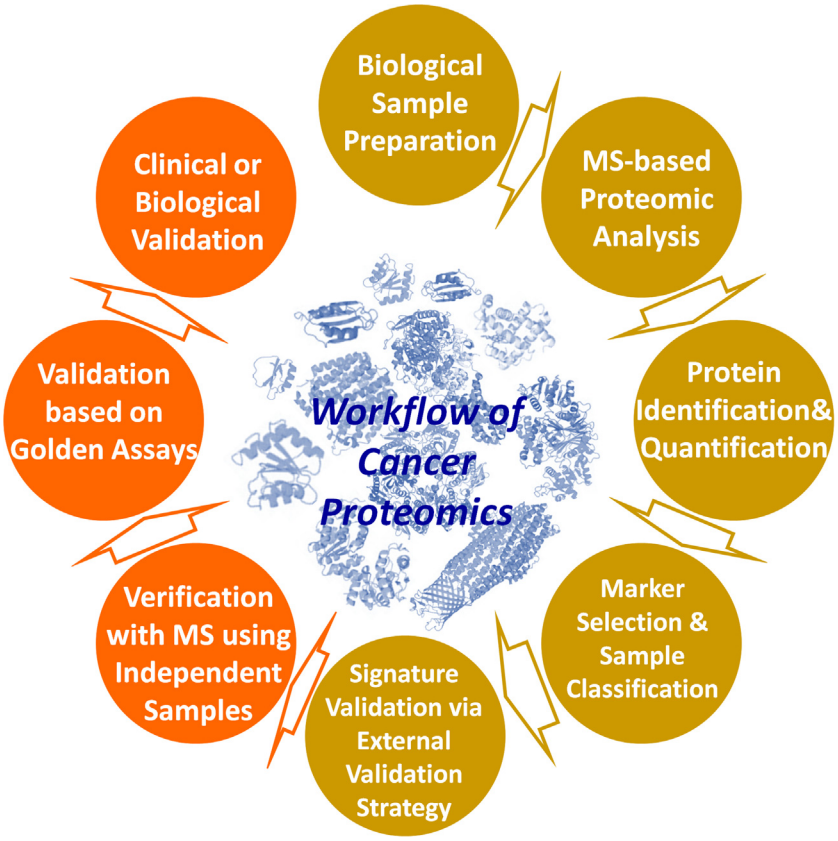

Fig. 1. The workflow of cancer biomarker discovery in quantitative proteomics study.

biological samples [33,34], It is also a powerful technique for identifying potential proteomic biomarkers, facilitating the discovery of anticancer drug targets and providing new sight into mechanism underlying complicated cancers [35–37,17]. The workflow of cancer biomarker discovery based on quantitative proteomics is illustrated in Fig. 1. In the proteomics study, it is critical for assessing the performance of feature selection methods using the external validations strategy [38]. The external validation implies the presence of a test dataset of samples that has not been employed for constructing the model, and could be considered the suitable strategy for avoiding overfitting problem [38].

Due to the sparsity of features in a big proteomic data, feature selection methods are applied to identify significant proteins/peptides (or features) between distinct groups, which is a crucial step for cancer classification, diagnosis and prognosis in the comparative proteomics study [38–42]. Currently, diverse feature selections have been developed and applied to the analyzed proteomic data in various cancer-associated studies [41]. However, due to the lack of robustness of feature selection methods, the robustness and consistency of the biomarker sets discovered in most cases is ambiguous [43,44].

This review describes recent computational advances made in the field of cancer proteomics from biomarker candidates' identification perspectives. A variety of popular feature selection techniques applied to cancer proteomic data are overviewed with critical assessments of both challenges and potentials. The main aim of this review is to make practitioners aware of the benefits and the necessity of applying feature selection techniques.

2. Considerations concerning study design

Tumor heterogeneity, design of study, sample size sand selection of the reference sample should be considered for discovering tumor biomarker in proteomic study [45–47]. For example, the use of normal/healthy controls as a reference group also needs to be made cautiously [47]. The inter-individual heterogeneity among tumors could affect proteins biomarker expression [48] and could result in inappropriate "normal controls" samples employed for tumor biomarker discovery [48]. Moreover, a relatively large numbers of samples employed could provide the biomarker validation in study design [49].

3. Feature selection for tumor marker identification

Feature selection techniques have become quite popular as well as required in tumor marker identification [50–53]. The application of feature selection techniques not only facilitates the identification of optimal differential oncological protein features [54–56], but also improves cancer risk stratification and prediction [57,58]. The available feature selection techniques can be organized into three categories based on their theories of screening variables and classifying distinct groups: filter methods, wrapper methods and embedded approaches. Moreover, one of the major problems in proteomics dataset is the treating of missing values. The imputation methods widely applied in proteomics data included the background imputation, Bayesian principal component imputation, censored imputation, k-nearest neighbor imputation, local least squares imputation and singular value decomposition [59].

As shown in Fig. 2, a common taxonomy of feature selection methods in different categories is provided, and the corresponding advantages and disadvantages of these methods are discussed with examples of the most influential techniques. Due to the various types of feature selection techniques available for tumor marker identification (Table 1), it is a great challenge to identify the optimal approach for analyzing any tumor marker-related study. The extensive application of each feature selection methods in current proteomic study together with popular sample classification methods are shown in Fig. 3. These common feature selection techniques currently available for tumor marker identification are overviewed as follow.

3.1. Filter methods for tumor marker identification from proteomic data

Filter methods involve in choosing the significantly differential features based on discriminating metric (e.g. *p*-value) that are relatively independent of classification. This metric well reflects the quality of each feature in terms of its discriminative power. The features would be remained when the metric values were within a specific criterion, and features beyond the thresholding condition would be eliminated. However, it should be noted that the feature selection based on p-value could be misleading when the difference is not based on regulation, but on the mere capacity of a cell to produce enough of the protein. Moreover, multicollinearity

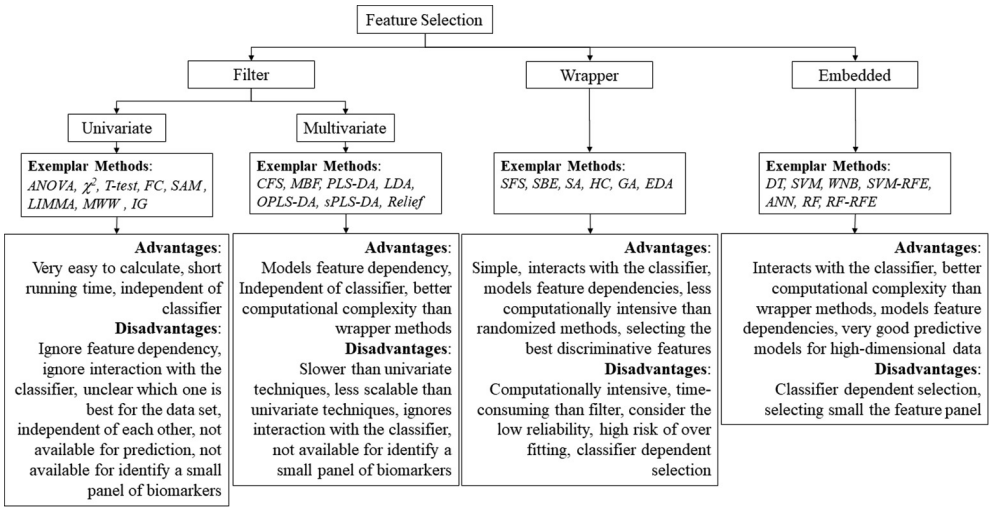

Fig. 2. Classification of available feature selection algorithms in cancer proteomic study.

Table 1
Different feature selection methods for identifying the markers in proteomics research.

	Methods	Extensive application of each feature selection methods in current proteomic research
Filter-based feature selection methods		
1	Fold change	Used to identify the protein markers for early screening detection and monitoring invasive breast cancer progression (*PLoS One*. 10:e0141876, 2015).
2	Entropy-based Filters	Applied to discover biomarkers from serum peptide profiling based on proteomic techniques (*Nat Protoc*. 2:588–602, 2007)
3	Analysis of variance	Used to discover the biomarkers associated with the orthostatic hypotension based on the proteomic study (*Hypertension*. 71:465–72, 2018)
4	Chi-square	Applied to reveal a prognostic signature in oral cancer in combining discovery and targeted proteomics (*Nat Commun*. 9:3598, 2018)
5	Student's t-distribution	Performed to reveal the differential proteins between samples with and without primary dysmenorrhea (*F1000Res*. 7:59, 2018)
6	Significance Analysis of Microarrays	Used to analyze significance between two biological samples in quantitative proteomic study (*BMC Bioinformatics*. 9:187, 2008)
7	Linear Models for Microarray Data	Applied to find significantly affected proteins by cathepsin A via proteomic profiling (*J Proteome Res*. 15:3188–95, 2016)
8	Linear discriminant analysis	Used to find circulating lipids and coagulation cascade in septic shock progression based on proteomics data (*Sci Rep*. 8:6681, 2018)
9	Mann–Whitney–Wilcoxon test	Used to identify the candidate biomarkers of prostate cancer in proteomic study (*BMC Bioinformatics*. 16:169, 2015)
10	Correlation-based feature selection	Used to screen the features for investigating the periplasmic expression of soluble proteins in Escherichia coli (*Sci Rep*. 6:21844, 2016)
11	Markov blanket filter	Used to detect proteomic biomarker in the recurrent ovarian cancer study from high-resolution mass spectrometry data (*IEEE Trans Inf Technol Biomed*. 13:195–206, 2009)
12	Partial Least Square Discriminant Analysis	Applied to select the discriminative proteins between the medullary sponge kidney and idiopathic calcium nephrolithiasis patients (*Int J Mol Sci*. 20:5517, 2019)
13	Orthogonal Partial Least Squares Discriminant Analysis	Used to identify candidate biomarkers significantly discriminated cholangiocarcinoma from normal and periductal fibrosis patients (*PLoS One*. 14:e0221024, 2019)
14	Sparse Partial Least Squares Discriminant Analysis	Applied to discovery the protein biomarkers in hypertrophic cardiomyopathy based on proteomics profiling (*J Cardiovasc Transl Res*. 12:569–79, 2019)
15	Discriminant Analysis of Principal Component	Used to find conditions for the subsequent identification of biomarkers and stress proteins (*PLoS One*. 11:e0165504, 2016)
Wrapper-based feature selection methods		
1	Sequential forward selection	Used for protein mass spectrometry in the disease diagnosis and biomarker identification (*BMC Bioinformatics*. 6:68, 2005)
2	Sequential backward elimination	Used to select k-spaced amino acid pairs to identify the pupylated proteins and pupylation sites based on proteomic data (*J Theor Biol*. 336:11–7, 2013)
3	Simulated annealing	Applied to find a biomarker in prostate proteomic pattern based on prostate protein mass spectrometry data (*CIBCB*, 195–200, 2008)
4	Hill climbings	Employed to tumor detection and discriminate tumor samples from nontumor ones in proteomic study (*Artif Intell Med*. 32:71–83, 2004)
5	Genetic algorithm	Used to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease (*Lancet*. 359:572–7, 2002)
6	Estimation of distribution algorithm	Applied to identify the prognosis biomarker of ovarian cancer in metabolomic study (*Metabolomics*, 7:614–22 2011)
Embedded feature selection methods		
1	Decision Tree	Used to automatically determine proteomic biomarkers and predictive models in the diagnosis of rheumatoid arthritis and inflammatory bowel diseases (*Bioinformatics*. 21:3138–45, 2005)
2	Support Vector Machine	Used to find proteins which differ between the breast cancer subtypes based on proteomic profiles (*Nat Commun*. 7:10259, 2016)
3	Weighted naïve Bayes	Used to discover serum protein biomarkers for the diagnosis of active pulmonary tuberculosis based on proteomic profiling (*J Clin Microbiol*. 55:3057–71, 2017)
4	SVM Recursive Feature Elimination	Used to identify the biomarkers for predicting the early recurrence of ovarian cancer based on serum proteomic profiling (*Genome Inform*. 16:195–204, 2005)
5	Artificial Neural Networks	Applied to identify serum signatures for detecting hepatocellular carcinoma and its subtypes (*Clin Chem*. 49:752–60, 2003)
6	Random Forest	Used to identify the protein biomarkers in non-small cell lung cancer for early stage asymptomatic patients (*Cancer Genomics Proteomics*. 16:229–44, 2019)
7	RF Recursive Feature Elimination	Applied to preliminarily screening the differential proteins for discovering biomarkers of hepatocellular carcinoma (*J Proteomics*. 225:103780, 2020)

implies a strong correlation between features that affect the target vectors simultaneously. The multicollinearity problem could lead to instability of some feature selection method [60]. The type of filter methods is relative of great simplicity compared to other wrapper and embedded methods. In particular, these filter methods can be organized into two categories based on the number of variables analyzed: (1) univariate and (2) multivariate methods. The former type refers to methods where only one variable is used for filter analysis, which includes Fold Change (FC), Entropy-based Filters, Analysis of Variance (ANOVA), Chi-square (χ^2), Euclidian Distance, *T*-test, Information Gain (IG), Significance Analysis of Microarrays (SAM), Linear Models for Microarray Data (LIMMA) and Mann–W hitney–Wilcoxon test (MWW test). The later type refers to methods where at least two variables are considered for building modes, which include Correlation-based Feature Selection (CFS), Markov Blanket Filter (MBF), Linear Discriminant Analysis (LDA), Partial Least Square Discriminant Analysis (PLS-DA), Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA), Sparse Partial Least Squares Discriminant Analysis (sPLSDA), Discriminant Analysis of Principal Component (DAPC) and Relief. As reported previously, the univariate methods could be used to evaluate a single biomarker with respect to recurrence and outcome, and multivariate methods were employed to develop a prognostic classification panel for disease recurrence [61]. The application of univariate and multivariate methods usually results in numerous significance features. Advantages of filter techniques are computationally simple

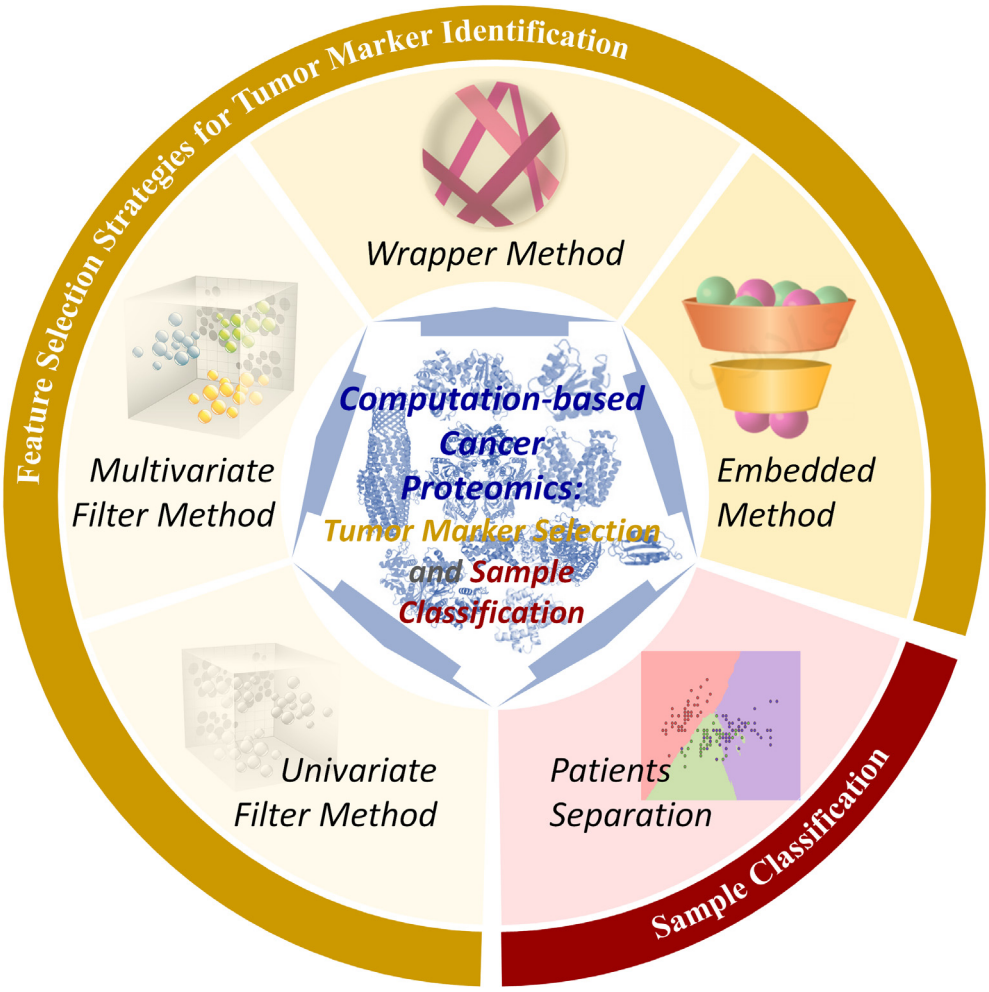

Fig. 3. Feature selections applied in cancer proteomics and popular sample classification methods.

and fast, and they are independent of the classification algorithm. A common disadvantage of filter methods is that they ignore the interaction with the classifier. Explanation and summary on each filter method are provided in Table 2. And the detailed descriptions and the corresponding applications of cancer-regulated studies are as follow:

3.1.1. Univariate filter methods

Fold Change (FC) is to compare the absolute value change between means of two groups, and it is calculated as the ratio or log of the ratio of the mean metabolite levels between two groups. A threshold needs to be defined, if fold change value exceeds this threshold, the variable will be reported as significant. FC has been widely applied in multiple omics analysis to predict survival of patients with metastatic colorectal cancer [62], and detect changes in the metabolomics profiles of men before and after androgen deprivation therapy for prostate cancer [63].

Entropy-based Filters are filter-based feature ranking techniques which include three classes: Information Gain (IG), Gain Ratio (GR) and Symmetrical Uncertainty (SU). IG selects features based on contribution of information related to the class variable without considering feature interactions. GR is non-symmetrical measure that is introduced to compensate for the bias of the IG, and SU criterion compensates for the inherent bias of IG [64]. Entropy-based Filters have been applied to identify the features related to glioma subtypes [65] and discover biomarkers in Arabidopsis thaliana [66].

A collection of statistical models and their associated estimation procedures was developed by Ronald Fisher called Analysis of Variance (ANOVA) [67]. This method mainly focuses on analyzing the differences among different groups average or variance values in a specific protein [68,69]. When comparing with multiple groups average values for statistical significance, it is more conservative than the traditional multiple two-sample *t*-tests method, and is therefore applicative for an extensive range of practical problems [70]. The ANOVA has been applied to investigate pro-inflammatory cytokines and liver functional markers [71] and explore the cellular resistance through quantitative proteomic analysis [72].

When a hypothetical feature is actual independent of the class value, Chi-square (χ^2) becomes a popular statistical test for weighting divergence distribution [73]. It is known that the χ^2 test is used to judge the independence of two events and its behavior is erratic for very small counts of cases [74]. Using the χ^2 statistic for feature selection is similar to importing a hypothesis testing about the distribution of classes [75]. Chi-square has been applied to study the incidence of injury to parathyroid glands during surgery for papillary thyroid carcinoma [76]. Main function of *t*-test is to judge whether the average of the two groups is statistically different [77]. *T*-test is suitable for samples with a normal distribution [78]. Commonly, the log transformation was needed for obtaining a more symmetric distribution prior to *t*-test analysis. Moreover, normalization is also a necessary step for proteomics analysis that aims to reduce systematic bias and make samples more compara-

Table 2
Fifteen filter feature selection methods available and popular in cancer proteomic study.

Methods	Abbr.	Packages in R (*Function*)	Brief Descriptions	Reference
Fold change	FC	metabolomics (*FoldChange*)	The FC is a popular and simple statistical method for the absolute value change between means of two groups. And it is often applied together with other parametric/non-parametric methods	*Clin Cancer Res.* 18:3677–85, 2012
Entropy-based Filters	Entropy	FSelector (*entropy. based*)	The Entropy filter method is similar to the Chi-square, which screens variables reraedless of possible interactions. And its performance depended on the characteristic s of studied data.	*Comput Chem Eng.* 22:613–626, 1998.s
Analysis of variance	ANOVA	ANOVA.TFNs (*fanova*)	The ANOVA is a powerful statistical method of parametric/non-parametric and linear. And focus on analyzing the differences among multiple groups average values in a specific protein.	*Br J Math Stat Psychol.* 68:23–42, 2015
Chi-square	χ^2	stats (*chisq. test*)	The χ^2 is a popular statistical test of non-parametric, and difficult interpreting when the present of a large number of categories in the variables.	*Am J Orthod Dentofacial Orthop.* 150:1066–1067, 2016
Student's t-distribution	*T*-test	stats (*t.test*)	The *T*-test is parametric method to judge whether the average of the two groups is statistically different.	*Br J Educ Psychol.* 76:663–75, 2006
Significance Analysis of Microarrays	SAM	samr (*sam*)	The SAM is non-parametric and linear statistical technique. And rely on samples when carries out variable specific *t*-tests and ignore single variable	*Proc Natl Acad Sci USA.* 98:5116–21, 2001
Linear Models for Microarray Data	LIMMA	limma (lmFit)	The LIMMA is an advanced statistical method based on parametric and linear models. And analyze the reverse-phase protein array data	*Nucleic Acids Res.* 43:e47, 2015
Linear discriminant analysis	LDA	MASS (*lda*)	The LDA is statistical algorithm based on the linear theory existing among different variables. And it is well suitable for study the small number of features.	*Methods Mol Biol.* 1362:175–84, 2016
Mann–Whitney–Wilcoxon test	MWW test	stats (*wilcox. test*)	The MWW test is a nonparametric test of the null hypothesis. And it is well suitable when the assumptions of the *t* test or the data is ordered are not met.	*J Wound Ostomy Continence Nurs.* 24:12, 1997
Correlation-based feature selection	CFS	FSelector (*cfs*)	The CFS is a parametric and linear algorithm that have a heuristic search strategy. Its goal is to finalize a subset of the characteristics of these groups that are less correlated but highly.	*Information Sciences.* 282:111–135, 2014
Markov blanket filter	MBF	MXM (*mmmb*)	The MBF is a based on discretized features feature selection statistical method, which can be applied for obtaining good classification performance with the small variables sets.	*Med Phys.* 42:2421–30, 2015
Partial Least Square Discriminant Analysis	PLS-DA	ropls (*opls*) caret (*plsda*)	The PLS-DA is supervised multivariate statistical approach and belongs to type of linear two-class classifier. And is can maximize interval between predefined classes.	*Chemometrics&Intelligent Laboratory Systems.* 58.2:109–130, 2001
Orthogonal Partial Least Squares Discriminant Analysis	OPLS-DA	ropls (*opls*)	The OPLS-DA is an advanced multivariate statistical method for one or more classes problems. And it is well suitable for classification of proteomic dataset when the variables often existed multi-collinear and noisy problems.	*J. Chemometrics.* 20:341–351, 2006
Sparse Partial Least Squares Discriminant Analysis	sPLSDA	ropls (*opls*) mixOmics (*splsda*)	The sPLS-DA is an advanced multivariate statistical algorithm for screening the variables and may miss out on variables between small sample sizes. The advantages of sPLS-DA are that variable selection and modeling are allowed in one step and can address multiclass problems.	*BMC Bioinformatics.* 12:253, 2011
Discriminant Analysis of Principal Components	DAPC	adegenet (*dapc*)	The DAPC is nonparametric and linear method. DAPC retains all property of DA without being dragged down by its restrictions.	*BMC Genet.* 11:94, 2010

ble [79–84]. As reported, the variance stabilization normalization (VSN) method has a built-in transformation and could perform well in the statistical analysis [59,85]. However, *t*-test with multiple testing correction might not be appropriate when studied sample size is too small (less than 6) [16]. The Multiple testing correction is a necessary step to manage the problem of false positive when performing multiple statistical tests in proteomics study [38]. Popularly applied multiple testing correction methods included the Bonferroni correction, Holm correction and Benjamini-Hochberg correction. The *t*-test has been applied to distinguish prostate cancer from normal and benign conditions [86].

In 2001, Virginia Tusher et al. developed a statistical technique called Significance Analysis of Microarrays (SAM) for judging whether statistical significance of gene expression changes exists [87]. SAM can assign a grade to every protein according to the distinction in expression associative with the standard deviation and then give a permutation based false discovery rate (FDR) estimate [87,88]. The advantage of this method is non-parametric statistics which uses repeated permutations of the data to judge whether the expression of protein is significant related to the response that is based on experimental conditions regardless of the distribution of individual proteins [87]. SAM has been applied to identify specific serum proteomic features associated with hepatocellular carcinoma [89].

A package called Linear Models for Microarray Data (LIMMA) focuses on differential expression gene analysis of file generated by simple or complex microarray experiments [90]. Matching a linear model to the gene or protein expression data is a central idea of this method. The functions of input and normalization can be performed by LIMMA when two color microarray data need to be processed [91]. Meanwhile, LIMMA can conjunct with the affyPLM or affy packages for Affymetrix data. It has been used in dopaminergic neurons function analysis for mouse and human fibroblasts [92] and analysis of reverse-phase protein array data [93]. LIMMA has been applied to investigate the influence of Cathepsin A on the cardiac proteome in a mouse model of cardiomyocyte-specific human Cathepsin A overexpression [94].

A non-parametric alternative test of the null hypothesis called the Mann–Whitney–Wilcoxon test (MWW test) aims at comparing two mean values from the same sample and testing if two sample values are equally distributed [95,96]. In general, the MWW test is used when the assumptions of the *t* test is not met [97]. By comparing the distribution of results of two groups with outliers in the data, MWW test is widely used to observe the treatment effect

[98]. As reported, its performance would be dramatically improved with sample size increasing to a certain degree (more than 12) when compared with the other methods (e.g. PCDA) [16,99]. The MWW test has been applied to identify candidate biomarkers of prostate cancer in proteomic study [100]. This method is implemented based on the linear theory existing among features. The LDA has been applied to identify a set of physicochemical properties for facilitating the classification of metal-based colloids [101].

3.1.2. Multivariate filter methods

Feature subsets among the features can be selected by Correlation-based Feature Selection (CFS) relying on the degree of redundancy [102]. The evaluator's goal is to finalize a subset of the characteristics of these groups that are less correlated but highly related to the class [103]. In particular, the subset evaluators search iteratively and add features using a numeric measure, for instance, conditional entropy [104]. As a multivariate filter, the CFS considers the relationship between features not noticed by non-univariate filters [103]. CFS is applied to select first-rate feature subset and is bonded with search methods, for example, bi-directional search, forward selection, genetic search and best-first search [105]. CFS has been used in the subtype discovery of pediatric acute lymphoblastic leukemia by cancer microarray file [106]. Based on discretized features, a feature selection method named Markov Blanket Filter (MBF) is raised [107]. As a critical evaluation of MBF, contradictory and counterintuitive nature leads to undesirable properties on the small sample size applications which have function of classifying for microarray gene expression data [108]. Linear Discriminant Analysis (LDA) is applied to identify differences or discriminant features among different samples groups and pattern classification [109], which is quite suitable for studying the small number of features [110].

Swedish statistician Herman O. A. Wold and his son Svante Wold proposed and developed Partial Least Square Discriminant Analysis (PLS-DA) [111]. This method belongs to linear two-class classifier and maximize interval between predefined classes [112]. The PLS-DA method will not produce the most accurate decision boundary once the sample sizes are unequal [111]. This method is now usually applied by chemometricians who incorporate it into most pack-ages and is therefore cited in metabolomics papers [113]. The PLS-DA combined other univariate methods has been applied to select the discriminative proteins between the medullary sponge kidney and idiopathic calcium nephrolithiasis patients [114].

A supervised multiple regression analysis called Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) can be applied to identify the difference between different data sets of X and Y [115]. The OPLS methods is an extended Partial Least Squares (PLS) method that combines an orthogonal signal correction filter to discriminate data changes from predictive quantitative response predictions that are useful from orthogonal predictions [116]. This method is a powerful tool to analyze qualitative data structures and predicted results are similar to the results of standard PLS-DA [117]. OPLS-DA has a primary advantage as for interpretation of the models due to its predictability [118]. The OPLS-DA has been applied to identify candidate biomarkers significantly discriminated cholangiocarcinoma from normal and periductal fibrosis patients [119]. Moreover, similar to the partial least squares discriminant analysis, the Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) aims to achieve discrimination analysis based on PLS regression [120]. When adding a Lasso penalization to select variables, sPLS-DA can select variable and reduce dimension simultaneously [120]. sPLS-DA may tend to miss out on variables that only distinguish between small sample sizes, assuming effect size [121]. The advantage of sPLS-DA is that variable selection and modeling are allowed in one step, and interpretability is improved through valuable graphical output [122].

In 2010, Jombart and colleagues proposed Discriminant Analysis of Principal Components (DAPC) for inferring the number of genetically related individuals [123]. DAPC and previously mentioned PLS-DA choose small-size feature sets precisely but miss many true positive features relevant to the spiked proteins [16]. This multivariate statistical approach divides variance of the sample into between-group and within-group to maximize discrimination between groups. DAPC first converts the data using Principal Component Analysis (PCA) and then uses Discriminant Analysis (DA) to identify the cluster [123]. As a new methodological approach, DAPC retains all property of DA without being dragged down by its restrictions [123]. A major advantage of DAPC in the Bayesian Cluster approach is the ability to generate graphs that infer kinship between clusters. DAPC can be used to decompose potential structures in more complex demographic models [124].

3.2. Wrapper methods for tumor marker identification from proteomic data

The wrapper method is one of the most popular feature selection methods proposed by Kohavi and John in 1997 [125]. In the wrapper method, a search for an optimal set of features is made using the induction algorithm as a black box [126]. It provides a simple and powerful technique to solve the challenge of feature selection. The wrapper methods look for the best subset of features based on their predictive power. Wrapper methods are often criticized because they train various new subsets into corresponding models, and need a large amount of computation, but it is not necessarily the case. Researchers can design an effective search strategy. Using this strategy does not necessarily mean sacrificing predictive performance. But it usually provides the best performing feature set for a particular type of model. Generally, different wrapper methods are used to select the best subset. The commonly used methods are Genetic Algorithms (GA), Hill Climbing (HC), Sequential Backward Elimination (SBE), Estimation of Distribution Algorithm (EDA), Simulated Annealing (SA), and Sequential Forward Selection (SFS). Explanation and summary on each wrapper method are provided in Table 3. And the detailed descriptions and the corresponding application of cancer-regulated studies are as follow:

A valid and commonly applicable function minimization method is the Genetic Algorithm (GA) [127]. This method achieves the results of global optimization based on the genetic theory of "survival of the fittest". Unlike traditional search methods, GA begins with a set of finite-length encoded alphabet strings not a real set of parameters [128]. Selection, crossover, and mutation are operators of GA [129]. Through the selection step, a number of individuals which have strong adaptability are found out from the group, and then the individual copy number is determined based on the selection method [130]. Generally speaking, it is necessary to select the suitable crossover and mutation probabilities based on practical problems [131]. GA had been used to maximize functional value in chronic dialysis patients [132]. Moreover, the most famous local search algorithm may be Hill Climbing (HC) [133]. The main idea of climbing is divided into three steps: first, randomly generate a state, and then find the best assessment value to transfer to the neighbor, and finally if the third step has reached a rigid local minimum, restart other randomly created states. These steps need to be repeated until the solution or a local optimum is found [134,135]. The HC has been applied to predict the protein structure prediction combining with genetic algorithm [136].

The Sequential Backward Elimination (SBE), also called SBS (Sequential Backward Selection), starts from the entire collection, discarding the feature x in turn minimizes the value of the objec-

Table 3
Six wrapper feature selection methods available and popular in cancer proteomic study.

Method	Abbr.	Packages in R (*Function*)	Brief Description	References
Sequential forward selection	SFS	Dprep (*sffs*)	SFS is the only method containing F core steps of linear parameter and is a bottom-up search technique. Many variants and applications are proposed based on SFS.	*Pattern recognition and signal processing* 41–60, 1978
Sequential backward elimination	SBE	MASS (stepAIC)	SBE is a linear and parameter algorithm that performs well due to its addressing of large subsets. SBE works in opposite direction.	*Pattern recognition and signal processing* 41–60, 1978
Simulated annealing	SA	GenSA (*GenSA*) SS	SA is a specific parameters and linear method of approximating the global optimization problems. It is strict with various conditions, but sometimes the cost-increasing can be acceptable.	*Soviet Phys Cryst.* 5:905–916, 1979
Hill climbing	HC	FSelector (*hill.climbing.search*)	HC is a local search linear and parameter algorithm, which does not need to find the global maximum. Combining with genetic algorithm, HC has been used to predict protein structure.	*Proteome Sci.* 1:S19, 2011
Genetic algorithm	GA	GA (ga)	GA is an effective and universally applicable function of parameters and linear/unlinear, which does not scale well with complexity.	*Mitochondrial DNA* 25:231–7, 2014
Estimation of distribution algorithm	EDA	Copulaedas (Copula)	EDA is a parameter approach and a widely used random optimization. It belongs to the class of evolutionary algorithms.	*IJCSE.* 3:1787–97, 2011

tive function J(Y-x-) [137,138]. These features are considered non-monotonic. Because SBS is addressing large subsets in most cases, it performs best while the first-rank feature subset exists many features [139]. The primary disadvantage of SBS is that it is not possible to re-evaluate the reusability of this feature when a feature is removed [140]. Sequential Forward Selection (SFS), a bottom-up search technique, works in opposite directions from SBE. SFS is first based on a null feature set and then gradually adds features picked by some assessment functions to minimize Mean Square Error (MSE) [141]. Sharma et al. proposed SFS [142] and attempted to conquer the shortcoming of traditional feature selection method whereby a weakly ranked protein that could conduce to classification accuracy with a suitable subset of proteins has eliminated from the selection [143]. In each iteration, the remaining available features that do not exist in the feature set are selected into the feature set [144]. Therefore, the extended feature set will generate a minimal classification deviation compared to the previously added features [145]. Despite this, SFS is widely applied because of its simplicity and the fact that it only contains F core steps, and many variants and applications are proposed based on the SFS.

Estimation of Distribution Algorithm (EDA), also named Probabilistic Model-building Genetic Algorithm (PMBGA), is a widely used random optimization method [146]. Firstly, EDA learns the explicit probability model which is a promising solution that is currently found, and then generates a new solution by sampling the created model [147]. EDA belongs to a class of evolutionary algorithm. Unlike most traditional evolutionary algorithms which use implicit distributions defined by one or more variant operators to generate new candidate solution, EDA uses Bayesian networks multivariable normal [148]. EDA is an excellent solution to a number of challenging problems [149]. EDA now has been applied to the HP model on a cubic lattice and predict the native structures of relatively small proteins therefore helps in drug design and proteomics [150–152]. Moreover, the Simulated Annealing (SA) is a specific method for a number of type optimization problems [153]. SA has now been applied to help identify HER2 and hormonal receptor expression states of breast cancer patients [154].

3.3. Embedded methods for tumor marker identification from proteomic data

Embedded methods are commonly used in feature selection in order to select the optimal subset of features which play a great role in the process of constructing the classifier [155]. A low minimum sample size of around 20 samples was suggested as sufficient to perform robust power analysis [156]. It is clear that sample imbalance has a great effect on identifying differential

Table 4
Seven embedded feature selection methods available and popular in cancer proteomic study.

Methods	Abbr.	Packages in R (*Function*)	Brief Descriptions	Reference
Decision Tree	DT	dtree (*pca*)	The DT is a supervised algorithm for classification and prediction. And it is suitable for nonlinear proteomic data.	*Machine Learning.* 4:161–86, 1989
Support Vector Machine	SVM	*e1071* (*svm*)	The SVM is a supervised learning model for classification and regression. And it is suitable any linear or nonlinear proteomic data, but also complex and time consuming for large scale proteomic data.	*Bioinformatics.* 21:47–56, 2005
Weighted naïve Bayes	WNB	CORElearn (*CoreModel*)	The WNB is a supervised learning method based on the Bayes theory. And it is only suitable for the proteomic data, where all variables are independent of each other.	*Appl Environ Microbiol.* 73:5261–7, 2007
SVM Recursive Feature Elimination	SVM-RFE	MCRestimate (*varSel.svm.rfe*)	The SVM-RFE is popular supervised embedded method. And it is well suitable for addressing overfitting risk when the number of variables is high.	*Anal Chem.* 80: 7562–70, 2008
Artificial Neural Networks	ANN	*neuralnet* (*neuralnet*)	The ANN is supervised embedded method. And its performance on consistent depended on the characteristic of studied data.	*Nat Rev Urol.* 10:174–82, 2013
Random Forest	RF	randomForest (*randomFores*)	The RF is popular supervised embedded method. And it is well suitable for any proteomic dataset for identifying the small variables set.	*Isprs Journal of Photogrammetry & Remote Sensing.* 67:93–104, 2012
RF Recursive Feature Elimination	RF-RFE	varSelRF (*varSelRF*)	The RF-RFE is popular supervised embedded method. And it is well suitable for addressing the non-linear or linear variables of proteomic dataset.	Metabolomics. 18: 3677–85, 2012

features. For example, the imbalance sample in the dataset could lead to model overfitting or underfitting. To deal with the overfitting problem, the dataset could be randomly divided into training (60%), validation (20%), and test (20%) data sets, or training (70%) and test (30%) data sets in case validation is not required [157].

In this section, six methods are listed as follows: Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), Bayesian Classifier and Artificial Neural Network (ANN), Random Forest with Recursive Feature Elimination (RF-RFE) and Support Vector Machine with Recursive Feature Elimination (SVM-RFE) [84]. Explanation on each method is provided in Table 4, and detailed descriptions and the corresponding application of cancer-regulated studies are as follow:

Decision Tree learner is a common decision support tool that resembles a tree structure in which each non-leaf node represents a test on a specific characteristic, each branch means the result of the test, and each leaf node denotes a class label [158]. Through an iterative process, the Decision Tree learns from a series of training examples, selects an attribute and then splits a given set of examples based on the values of that attribute [159]. Advantages of the method can be summed up in two aspects. First, the construction of the classifier does not require much of professional knowledge and its reliability in diagnosis can be verified through both expert knowledge and testing data [160]. Besides, its inductive classification step is simple and fast [159].

Random Forest developed by Leo Breiman is an algorithm that integrates multiple trees through the idea of integrated learning [161]. Its basic unit is Decision Tree which takes the advantage of a bootstrap sample of the data [162]. The method is excellent in accuracy [163] and generates an internal unbiased estimate of the generalization error as the forest building progresses. Moreover, when comparing with SVM, it performs better in the aspect of classification tasks [164]. Random Forest with Recursive Feature Elimination (RF-RFE) conducts a recursive backward feature elimination procedure. It begins with all the features. In each iteration, a Random Forest is constructed to measure the features' importance and the feature with least importance is removed. This procedure is repeated until there is no feature left. Finally, the features are ranked according to the deleted sequence, and the top ranked feature is the last deleted one. RF-RFE has been applied to reveal the deregulation of fatty acids metabolism in hepatocellular carcinoma and chronic liver [165,166], and investigate hepatitis C by using metabolomics as a tool [63].

Support Vector Machine (SVM) is a kind of learning machines constructed according to structural risk minimization principle on the basis of Vector Machine theory, which hopes to find a segmentation surface that can separate the data points apart [167]. One of the challenges of this method is how to choose the proper kernel function [168]. When it comes to the advantages of SVM, it performs well in dealing with high-dimensional data sets with very few training examples [169]. Moreover, as family of simple "probabilistic classifiers" based on Bayes theory, the Bayesian Classifier can be applied to predict the probability that a given instance pertains to a class [170]. It assumes that all the attributes are independent of each other [171]. In theory, the naive Bayesian classifier has the minimum error rate when compared to other classification methods. However, this is not always the case in practice due to the above assumption. Even so, the Naive Bayesian classifier has the high precision and speed in the application of large database [172].

Support Vector Machine with Recursive Feature Elimination (SVM-RFE) is an embedded approach that uses the norm of the weights w to rank the variables. Initially, all data is taken, and a classifier is computed. The norm of w is then computed for each of the features and the feature with the smallest norm is eliminated. This process is repeated until all features are ranked. For selecting features relevant to the spiked compounds, linear SVM-RFE performs poorly even if the classification error is relatively low [16]. SVM-RFE has been applied to determine differences between healthy subjects and patients suffering from Streptococcus pneumoniae [173], and interrogate the serum metabolome of early-stage ovarian cancer patients and age-matched control women [174].

Artificial Neural Networks (ANN) is an algorithm mathematical that simulates the behavior characteristics of biological neural networks, including their structure and functionalities [175]. The basic unit of every artificial neural network is artificial neuron which is a simple mathematical function [176]. This kind of network relies on the complexity of the system and can process information by adjusting the weight of the interconnection between a large number of nodes (neurons) within the system [177]. Requiring less formal statistical training, ANN performs well in detecting the complex non-linear relationship between dependent and independent variables [178].

In addition to the above feature selection methods, deep learning is also a type of machine learning approaches which enables to identify the highly complex patterns in omics data [179–181]. Common deep learning methods included the convolutional neural networks, recurrent neural networks and deep neural networks [182]. Untill now, deep learning has been widely applied in cancer proteomics [183]. For example, the deep learning approach has been applied for the diagnosis of pancreatic cancer in proteomic data [183].

4. Sample clustering and visualization

Clustering or unsupervised modeling is a powerful technique for discovering the molecular classification of patient's cancer tissues [184] and identifying subtype specific characterization (e.g. subtype specific cancer therapeutic targets) [185,186]. The underlying theories of clustering technique are considering the similarities among different samples based on high dimension proteins or peptides, which can provide information of the same or differential patterns on protein expression level, aiming at clustering those biologically similar samples together [187]. Therefore, the unsupervised approaches are widely applied to characterize and analyze the complex proteomic dataset. Currently, there are five different approaches for sample separation of cancer proteomic dataset, including three clustering methods and two popular reducing dimension methods [188]. In cluster analysis, Hierarchical Cluster Analysis (HCA), Self-organizing Map (SOM) and K-means Clustering are three prominent representatives in the analysis of proteomic data [189]. Principal Component Analysis (PCA) and the state of art stochastic neighbor embedding analysis based on Student t Distribution (t-SNE) [190,184] are the most commonly used unsupervised approaches of reduction dimensions in cancer proteomic studies [188]. Explanation on these methods is provided in Table 5. Detailed descriptions and the corresponding application of cancer-regulated studies are as follow.

The Principal Component Analysis (PCA) aims at interpreting the first few principal components that often explain thousands of the variables, and is often applied strategy of reducing dimension for high dimensional cancer proteomic data [191]. Particularly, a multivariate proteomic dataset can first be transformed based on the linear relationship between different proteins or peptides [188], and then determine the most significant Principal Components (PCs) according to the obtained variances [191]. In PCA analysis, the scatterplots (e.g. scores plot and loadings plot) are often used for visualizing correlation of the biological samples or the importance of proteins/peptides to discriminate those differential samples, which are often visualized based on two axes (two

Table 5
Five methods for sample clustering and dimension reduction available in cancer proteomics.

Methods	Abbr.	Packages in R (*Function*)	Brief Descriptions	Reference
Principal Component Analysis	PCA	ropls (*pca*)	The PCA is unsupervised, non-parametric and linear method of reduce dimension method. And Unsuitable the dataset when the proteins or peptides were nonlinear relationship.	*Clin Biochem.* 56:55–61, 2016
t-Distributed Stochastic Neighbor Embedding	t-SNE	Rtsne (*Rtsne*)	The t-SNE is unsupervised, non-parametric and non-linear method of reduce dimension method. And Unsuitable the dataset when the proteins or peptides were linear relationship.	*J Mach Learn Res.* 9:25S79–605, 2008
Self-Organizing Map	SOM	Som (*som*)	The SOM is unsupervised, non-parametric and non-linear method of clustering. And suitable for any proteomic the dataset.	*IEEE Trans Neural Netw.* 11(3):574–85, 2000
Hierarchical Cluster Analysis	HCA	Cluster (*agnes*)	The HCA is unsupervised, non-parametric and non-linear method of clustering. And suitable the any datasets.	*Front Bioeng Biotechnols.* 3:23. 2015
K-Means Clustering	K-Means	skmeans (*skmeans*)	The K-Means is unsupervised, non-parametric and non-linear method of clustering. And Unsuitable the dataset when the number of clusters is not known.	*Proteomics.* 16:1613–21, 2016

principal components) or three axes (three principal components). The PCA scores plot demonstrates the discriminative power between different sample groups (these points represent the corresponding samples). The PCA scores plots establish whether there are any intrinsic differences in the composition of samples. Moreover, the loading plots of proteome from PCA models display proteins or peptides positively correlated with score plots. From the loading plots, differential proteins or peptides between control and cancer subjects are identified [192]. In general, the more these variables are away from center points, the more significant differences they can be considered. The PCA has been widely applied in comparative proteomic analysis, for example liver cirrhosis and gastric cancer samples clustering and visualization [193]. The PCA has been also applied in metabolomics to produce global metabolite profiles by studying perturbations under different biological [194] and studying toxicity induced by drugs [195].

Compared with the principal component analysis, the Stochastic Neighbor Embedding (SNE) method is a recently proposed and much more effective on-linear dimensionality algorithm of reducing dimension, which visualizes high-dimensional data by describing the similarity of samples using a low dimension space. The technique is primarily derived from SNE [196], and has unique advantages of using symmetrized SNE with straightforward gradients and computing the similarity between two samples based t distribution. Thus, SNE can alleviate crowded area of the two or three-dimensional map, and be quite suitable for exploring the optimal global organizations of the data based on the early exaggeration trick. SNE map is widely applied for cancer proteomic studies, for example, uncovering subtypes of gastric cancer and metastasis status in primary breast cancer [197].

When the number of clusters is unknown, the most prevalent clustering techniques would be unsupervised Hierarchical Clustering Analysis (HCA). The HCA can perform the clustering of samples by a dendrogram or tree plot [198]. The proteins expression levels or the studied samples can be clustered in a dendrogram [199]. The correlation coefficient across samples can be estimated based on the specific distance (e.g. Euclidean, Manhattan or Maximum distances). For these samples with the highest correlation coefficients, they would be categorized into the same cluster. The obtained dendrograms can explain many biological issues in the context of cancer subtypes and development statuses [200]. HCA is widely applied for cancer proteomic studies to identify the co-expression patterns of differentially expressed proteins in hepatocellular carcinoma based on iTRAQ quantitative proteomics analysis [201].

The K-means Clustering is prevalently used in cancer proteomic data when the number of underlying clusters is known [202]. The K-means Clustering aims at performing the grouping based on the known the clusters number k, and categorizing n samples into the k classes [203]. These samples with the most similarity values can be considered the same class, and the similarities are computed using Euclidean Distances. Compared with the HCA, the method often needs additional check procedures to identify the optimal clusters number. The method has been widely applied for identifying the co-expression patterns in glioblastoma multiform based the protein expression level [204]. Self-organizing Map analysis (SOM) is another popular clustering method applied in cancer proteomic study, which aims at describing the similarity of studied samples based on low dimensional map space [205]. The advantages of SOM are that it doesn't need to input the known cluster number and these similarity samples exist topological relationships. In summary, SOM regards the most similarity samples as in a clustering. Thus, SOM provides a map, in which, similar samples are grouped together, and disparate samples can be separated into different groups. SOM has been widely applied in multiple omics study to visualize samples phenotypes as well as variables patterns [188], for example, revealing unique dynamic expression patterns of proteins among the different colorectal cancer status [199]. SOM has also been applied to metabolic study for visualization of differential metabolites associated with wound response of *Arabidopsis thaliana* [206].

5. Conclusions

Tumor marker selection and sample classification are key issues in current cancer proteomics study. The biomarker sets identified via feature selection methods were inconsistency in several publications. This article reviewed the popular feature selection methods currently applied in tumor maker selection and discussed several sample classification algorithms available for cancer proteomic. These advances made the MS-based proteomics technology more widely applied to identify diagnostic, prognostic, and therapeutic biomarkers for anticancer drug discovery. Moreover, the pathogenetic process of a specific disease could be comprehensively and better understood using a panel of biomarkers, under this circumstance, the wrapper and embedded feature selection methods were suitable for selecting a small the feature panel in the tumor marker identifications.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Authors' contributions

F.Z. conceived the idea and supervised the work. J.T., Y.W. and Y. L. performed the research and wrote the manuscript. J.F., Y.Z., Y.L., Z.X., Y.L. and Y.Q. provide the review of cancer proteomics. All authors read and approved the final manuscript.

Declaration of competing interests

The authors declare that they have no competing interests.

Acknowledgements

Funded by the National Key Research and Development Program of China (2018YFC0910500), National Natural Science Foundation of China (81872798 & U1909208), Fundamental Research Funds for Central University (2018QNA7023, 10611CDJXZ238826, 2018CDQYSG0007 & CDJZR14468801), Key R&D Program of Zhejiang Province (2020C03010), & Leading Talent of "Ten Thousand Plan" - National High-Level Talents Special Support Plan.

References

[1] Malvezzi M, Carioli G, Bertuccio P, Negri E, La Vecchia C. Relation between mortality trends of cardiovascular diseases and selected cancers in the European Union, in 1970–2017. Focus on cohort and period effects. Eur J Cancer 2018;103:341–55.
[2] Arora D, Chaudhary R, Singh A. System biology approach to identify potential receptor for targeting cancer and biomolecular interaction studies of indole [2,1-a]isoquinoline derivative as anticancerous drug candidate against it. Interdiscip Sci Comput Life Sci 2017;11:125–34.
[3] Reddy MM, Kar SS. Unconditional probability of dying and age-specific mortality rate because of major non-communicable diseases in India: time trends from 2001 to 2013. J Postgrad Med 2018;65:11.
[4] Guo Y, Wang H, Li Y, Song Y, Chen C, Liao Y, et al. Genome of *Helicobacter pylori* strain XZ274, an isolate from a tibetan patient with gastric cancer in China. J Bacteriol 2012;194:4146–7.
[5] Fu J, Tang J, Wang Y, Cui X, Yang Q, Hong J, et al. Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. Front Pharmacol 2018;9:681.
[6] Collins LG, Haines C, Perkel R, Enck RE. Lung cancer: diagnosis and management. Am Fam Physician 2007;75:56–63.
[7] Yilmaz S, Dursun M, Canoruç F, Yilmaz S, Dursun M, Canoruç F. A patient with gastric carcinoid tumor: treatment and surveillance options. Turk J Gastroenterol 2005;16:180–1.
[8] Zhang L, Ozao J, Warner R, Divino C. Review of the pathogenesis, diagnosis, and management of type I gastric carcinoid tumor. World J Surg 2011;35:1879–86.
[9] Zhang A, Sun H, Yan G, Wang P, Han Y, Wang X. Metabolomics in diagnosis and biomarker discovery of colorectal cancer. Cancer Lett 2014;345:17–20.
[10] Singh S, Singh DB, Singh A, Gautam B, Ram G, Dwivedi S, et al. An approach for identification of novel drug targets in *Streptococcus pyogenes* SF370 through pathway analysis. Interdiscip Sci 2016;8:388–94.
[11] Ahmad S, Navid A, Akhtar AS, Azam SS, Wadood A, Perez-Sanchez H. Subtractive genomics, molecular docking and molecular dynamics simulation revealed LpxC as a potential drug target against multi-drug resistant *Klebsiella pneumoniae*. Interdiscip Sci 2018:1–19.
[12] Li S, Li J, Ning L, Wang S, Niu Y, Jin N, et al. In silico identification of protein S-palmitoylation sites and their involvement in human inherited disease. J Chem Inf Model 2015;55:2015–25.
[13] Qu KY, Gao F, Guo F, Zou Q. Taxonomy dimension reduction for colorectal cancer prediction. Comput Biol Chem 2019;83:107160.
[14] Liao ZJ, Li DP, Wang XR, Li LS, Zou Q. Cancer diagnosis through isomir expression with machine learning method. Curr Bioinform 2018;13:57–63.
[15] Li YH, Li XX, Hong JJ, Wang YX, Fu JB, Yang H, et al. Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. Brief Bioinform 2020;21:649–62.
[16] Christin C, Hoefsloot HC, Smilde AK, Hoekman B, Suits F, Bischoff R, et al. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. Mol Cell Proteomics 2013;12:263–76.
[17] Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res 2018;46:D1121–7.
[18] Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics 2018;34:398–406.
[19] Ji J, Tang J, Xia K-J, Jiang R. LncRNA in tumorigenesis microenvironment. Curr Bioinform 2019;14:640–1.
[20] Alvarez-Chaver P, Otero-Estevez O, Paez de la Cadena M, Rodriguez-Berrocal FJ, Martinez-Zorzano VS. Proteomics for discovery of candidate colorectal cancer biomarkers. World J Gastroenterol 2014;20:3804–24.
[21] Cai Y, Wang N, Wu X, Zheng K, Li Y. Compensatory variances of drug-induced hepatitis B virus YMDD mutations. Springerplus 2016;5:1340.
[22] Kondo T. Inconvenient truth: cancer biomarker development by using proteomics. BBA 2014;1844:861–5.
[23] Chang R, Shoemaker R, Wang W. Systematic search for recipes to generate induced pluripotent stem cells. PLoS Comput Biol 2011;7:e1002300.
[24] Tiss A, Timms J, Menon U, Gammerman A, Cramer R. Proteomics approaches towards early detection and diagnosis of ovarian cancer. J ImmunoTher Cancer 2014;2:O5.
[25] Li G, Xiao Z, Liu J, Li C, Li F, Chen Z. Cancer: a proteomic disease. Sci China Life Sci 2011;54:403–8.
[26] Lin M, Li X, Guo H, Ji F, Ye L, Ma X, et al. Identification of bone metastasis-associated genes of gastric cancer by genome-wide transcriptional profiling. Curr Bioinform 2019;14:62–9.
[27] Tsuchiya N, Sawada Y, Endo I, Saito K, Uemura Y, Nakatsura T. Biomarkers for the early diagnosis of hepatocellular carcinoma. World J Gastroenterol 2015;21:10573–83.
[28] Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A et al. Recommendations for biomarker identification and qualification in clinical proteomics. Sci Transl Med. 2010;2:46ps2.
[29] Karimi P, Shahrokni A, Ranjbar MR. Implementation of proteomics for cancer research: past, present, and future. Asian Pac J Cancer Prev 2014;15:2433–8.
[30] Yang QX, Wang YX, Li FC, Zhang S, Luo YC, Li Y, et al. Identification of the gene signature reflecting schizophrenia's etiology by constructing artificial intelligence-based method of enhanced reproducibility. CNS Neurosci Ther 2019;25:1054–63.
[31] Tang J, Wang Y, Fu J, Zhou Y, Luo Y, Zhang Y, et al. A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies. Brief Bioinform 2020;21:1378–90.
[32] 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. Jpn J Clin Oncol 2013;43:103–9.
[33] Distler U, Kuharev J, Navarro P, Tenzer S. Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics. Nat Protoc 2016;11:795–812.
[34] Shen W, Li Y. A novel algorithm for detecting multiple covariance and clustering of biological sequences. Sci Rep 2016;6:30425.
[35] Zhu F, Li XX, Yang SY, Chen YZ. Clinical success of drug targets prospectively predicted by in silico study. Trends Pharmacol Sci 2018;39:229–31.
[36] Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res 2020;48:D1031–41.
[37] Yin J, Sun W, Li F, Hong J, Li X, Zhou Y, et al. VARIDT 1.0: variability of drug transporter database. Nucleic Acids Res 2020;48:D1042–50.
[38] Lualdi M, Fasano M. Statistical analysis of proteomics data: a review on feature selection. J Proteomics 2019;198:18–26.
[39] Goh WW, Wong L. Evaluating feature-selection stability in next-generation proteomics. J Bioinform Comput Biol 2016;14:1650029.
[40] Goh WW. Fuzzy-FishNET: a highly reproducible protein complex-based approach for feature selection in comparative proteomics. BMC Med Genomics 2016;9:67.
[41] Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic Acids Res 2015;43:W251–7.
[42] Hoekman B, Breitling R, Suits F, Bischoff R, Horvatovich P. msCompare: a framework for quantitative analysis of label-free LC-MS data for comparative candidate biomarker studies. Mol Cell Proteomics 2012;11(M111):015974.
[43] Spratt HM, Ju H. Statistical approaches to candidate biomarker panel selection. Adv Exp Med Biol 2016;919:463–92.
[44] Yang Q, Li B, Tang J, Cui X, Wang Y, Li X, et al. Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. Brief Bioinform 2020;21:1058–68.
[45] Panis C, Pizzatti L, Souza GF, Abdelhay E. Clinical proteomics in cancer: where we are. Cancer Lett 2016;382:231–9.
[46] Panis C, Pizzatti L, Abdelhay E. How can proteomics reach cancer biomarkers?. Curr Proteomics 2013;10:136–49.
[47] Ignjatovic V, Geyer PE, Palaniappan KK, Chaaban JE, Omenn GS, Baker MS, et al. Mass spectrometry-based plasma proteomics: considerations from sample collection to achieving translational data. J Proteome Res 2019;18:4085–97.
[48] Dirks RC, Burney HN, Anjanappa M, Sandusky GE, Hao Y, Liu Y, et al. Breast heterogeneity: obstacles to developing universal biomarkers of breast cancer initiation and progression. J Am Coll Surg 2020;231:85–96.
[49] Jimenez CR, Verheul HM. Mass spectrometry-based proteomics: from cancer biology to protein biomarkers, drug targets, and clinical applications. Am Soc Clin Oncol Educ Book 2014:e504–10.
[50] Liu H, Xu Y, Xiang J, Long L, Green S, Yang Z, et al. Targeting alpha-fetoprotein (AFP)-MHC complex with CAR T-cell therapy for liver cancer. Clin Cancer Res 2017;23:478–88.

[51] Louis E, Adriaensens P, Guedens W, Vanhove K, Vandeurzen K, Darquennes K, et al. Metabolic phenotyping of human blood plasma: a powerful tool to discriminate between cancer types?. Ann Oncol 2016;27:178–84.
[52] Duan L, Yobas L. Label-free multiplexed electrical detection of cancer markers on a microchip featuring an integrated fluidic diode nanopore array. ACS Nano 2018;12:7892–900.
[53] Butti MD, Chanfreau H, Martinez D, Garcia D, Lacunza E, Abba MC. BioPlat: a software for human cancer biomarker discovery. Bioinformatics 2014;30:1782–4.
[54] Zduniak K, Ziolkowski P, Ahlin C, Agrawal A, Agrawal S, Blomqvist C, et al. Nuclear osteopontin-c is a prognostic breast cancer marker. Br J Cancer 2015;112:729–38.
[55] Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE 2016;11:e0163962.
[56] Tang J, Mou M, Wang Y, Luo Y, Zhu F. MetaFS: performance assessment of biomarker discovery in metaproteomics. Brief Bioinform 2020. https://doi.org/10.1093/bib/bbaa105.
[57] Avgeris M, Stamati L, Kontos CK, Piatopoulou D, Marmarinos A, Xagorari M, et al. BCL2L12 improves risk stratification and prediction of BFM-chemotherapy response in childhood acute lymphoblastic leukemia. Clin Chem Lab Med 2018;56:2104–18.
[58] Li Y, Chen M, Cao H, Zhu Y, Zheng J, Zhou H. Extraordinary GU-rich single-strand RNA identified from SARS coronavirus contributes an excessive innate immune response. Microbes Infect 2013;15:88–95.
[59] Valikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. Brief Bioinform 2018;19:1–11.
[60] Katrutsa A, Strijov V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Syst Appl 2017;76:1–11.
[61] Rinewalt D, Shersher DD, Daly S, Fhied C, Basu S, Mahon B, et al. Development of a serum biomarker panel predicting recurrence in stage I non-small cell lung cancer patients. J Thorac Cardiovasc Surg 2012;144:1344–50.
[62] Bertini I, Cacciatore S, Jensen BV, Schou JV, Johansen JS, Kruhoffer M, et al. Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. Cancer Res 2012;72:356–64.
[63] Saylor PJ, Karoly ED, Smith MR. Prospective study of changes in the metabolomic profiles of men during their first three months of androgen deprivation therapy for prostate cancer. Clin Cancer Res 2012;18:3677–85.
[64] Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, et al. Challenges in microbial ecology: building predictive understanding of community function and dynamics. ISME J 2016;10:2557–68.
[65] Elkhaled A, Jalbert L, Constantin A, Yoshihara HA, Phillips JJ, Molinaro AM, et al. Characterization of metabolites in infiltrating gliomas using ex vivo (1)H high-resolution magic angle spinning spectroscopy. NMR Biomed 2014;27:578–93.
[66] Lundstedt T, Hedenström M, Soeria-Atmadja D, Hammerling U, Gabrielsson J, Olsson J, et al. Dynamic modelling of time series data in nutritional metabonomics - a powerful complement to randomized clinical trials in functional food studies. Chemometr Intel Lab 2010;104:112–20.
[67] Kempthorne O. The correlation between relatives on the supposition of mendelian inheritance. Sci T R So 1919;52:399–433.
[68] McHugh ML. Multiple comparison analysis testing in ANOVA. Biochem Med 2011;21:203–9.
[69] Pritchard CC, Hsu L, Delrow J, Nelson PS. Project normal: defining normal variance in mouse gene expression. Proc Natl Acad Sci U S A 2001;98:13266–71.
[70] Jafari P, Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inf Decis Making 2006;6:27.
[71] Kermanizadeh A, Pojana G, Gaiser BK, Birkedal R, Bilanicova D, Wallin H, et al. In vitro assessment of engineered nanomaterials using a hepatocyte cell line: cytotoxicity, pro-inflammatory cytokines and functional markers. Nanotoxicology 2013;7:301–13.
[72] Zhao M, Li H, Bu X, Lei C, Fang Q, Hu Z. Quantitative proteomic analysis of cellular resistance to the nanoparticle abraxane. ACS Nano 2015;9:10099–112.
[73] Koletsi D, Pandis N. The chi-square test for trend. Am J Orthod Dentofacial Orthop 2016;150:1066–7.
[74] McHugh ML. The chi-square test of independence. Biochem Med 2013;23:143–9.
[75] Zhang H, Li L, Luo C, Sun C, Chen Y, Dai Z, et al. Informative gene selection and direct classification of tumor based on Chi-square test of pairwise gene interactions. Biomed Res Int 2014;2014:589290.
[76] Deng W, Li H, Chen Y, Gao Y, Huang H, Lin S, et al. Clinical application of carbon nanoparticles in surgery for papillary thyroid carcinoma in young patients. Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi 2014;49:812–6.
[77] Wright DB. Comparing groups in a before-after design: when t test and ANCOVA produce different results. Br J Educ Psychol 2006;76:663–75.
[78] Cibrik DM, Warner RL, Kommareddi M, Song P, Luan FL, Johnson KJ. Identification of a protein signature in renal allograft rejection. Proteomics Clin Appl 2013;7:839–49.
[79] Chawade A, Alexandersson E, Levander F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. J Proteome Res 2014;13:3114–20.
[80] Yang Q, Wang Y, Zhang Y, Li F, Xia W, Zhou Y, et al. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. Nucleic Acids Res 2020;48:W436–48.
[81] Yang Q, Hong J, Li Y, Xue W, Li S, Yang H, et al. A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies. Brief Bioinform 2019. https://doi.org/10.1093/bib/bbz137.
[82] Tang J, Fu J, Wang Y, Li B, Li Y, Yang Q, et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. Brief Bioinform 2020;21:621–36.
[83] Li B, Tang J, Yang Q, Li S, Cui X, Li Y, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic Acids Res 2017;45:W162–70.
[84] Li B, Tang J, Yang Q, Cui X, Li S, Chen S, et al. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. Sci Rep 2016;6:38881.
[85] Tang J, Fu J, Wang Y, Luo Y, Yang Q, Li B, et al. Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. Mol Cell Proteomics 2019;18:1683–99.
[86] Huo Q, Litherland SA, Sullivan S, Hallquist H, Decker DA, Rivera-Ramirez I. Developing a nanoparticle test for prostate cancer scoring. J Transl Med 2012;10:44.
[87] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–21.
[88] Langley SR, Mayr M. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. J Proteomics 2015;129:83–92.
[89] Poon TC, Yip TT, Chan AT, Yip C, Yip V, Mok TS, et al. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. Clin Chem 2003;49:752–60.
[90] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
[91] Diboun I, Wernisch L, Orengo CA, Koltzenburg M. Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. BMC Genomics 2006;7:252.
[92] Caiazzo M, Dell'Anno MT, Dvoretskova E, Lazarevic D, Taverna S, Leo D, et al. Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. Nature 2011;476:224–7.
[93] Mannsperger HA, Gade S, Henjes F, Beissbarth T, Korf U. RPPanalyzer: analysis of reverse-phase protein array data. Bioinformatics 2010;26:2202–3.
[94] Petrera A, Kern U, Linz D, Gomez-Auli A, Hohl M, Gassenhuber J, et al. Proteomic profiling of cardiomyocyte-specific cathepsin a overexpression links cathepsin A to the oxidative stress response. J Proteome Res 2016;15:3188–95.
[95] Whitney J. Testing for differences with the nonparametric mann-whitney u test. J Wound Ostomy Continence Nurs 1997;24:12.
[96] Marx A, Backes C, Meese E, Lenhof HP, Keller A. EDISON-WMW: exact dynamic programing solution of the wilcoxon-mann-whitney test. Genomics Proteomics Bioinformatics 2016;14:55–61.
[97] Tang Y. Size and power estimation for the wilcoxon-mann-whitney test for ordered categorical data. Stat Med 2011;30:3461–70.
[98] Wu P, Han Y, Chen T, Tu XM. Causal inference for mann-whitney-wilcoxon rank sum and other nonparametric statistics. Stat Med 2014;33:1261–71.
[99] Li F, Zhou Y, Zhang X, Tang J, Yang Q, Zhang Y, et al. SSizer: determining the sample sufficiency for comparative biological study. J Mol Biol 2020;432:3411–21.
[100] Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through venn diagrams. BMC Bioinf 2015;16:169.
[101] Sayes CM, Smith PA, Ivanov IV. A framework for grouping nanoparticles based on their measurable characteristics. Int J Nanomedicine 2013;8:45–56.
[102] Hall M. Correlation-based feature selection for machine learning. Waikato university; 1998.
[103] Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A, Benitez JM, Herrera F. A review of microarray datasets and applied feature selection methods. Inform Sci 2014;282:111–35.
[104] Xu J, Sun L, Gao Y, Xu T. An ensemble feature selection technique for cancer recognition. Biomed Mater Eng 2014;24:1001–8.
[105] Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, et al. Gene selection from microarray data for cancer classification-a machine learning approach. Comput Biol Chem 2005;29:37–46.
[106] Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 2002;1:133–43.
[107] Koller D, Sahami M, editors. Toward optimal feature selection. Thirteenth international conference on international conference on machine learning; 1996.
[108] Lee S, Ybarra N, Jeyaseelan K, Faria S, Kopek N, Brisebois P, et al. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. Med Phys 2015;42:2421–30.
[109] Kuligowski J, Perez-Guaita D, Quintas G. Application of discriminant analysis and cross-validation on proteomics data. Methods Mol Biol 2016;1362:175–84.

[110] Shi Y, Dai DQ, Liu CC, Yan H. Sparse discriminant analysis for breast cancer biomarker identification and classification. Prog Nat Sci-Mater 2009;19:1635–41.
[111] Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometr Intel Lab 2001;58:109–30.
[112] Wong KH, Razmovski-Naumovski V, Li KM, Li GQ, Chan K. Differentiation of *Pueraria lobata* and *Pueraria thomsonii* using partial least square discriminant analysis (PLS-DA). J Pharm Biomed Anal 2013;84:5–13.
[113] Nguyen HT, Lee DK, Lee WJ, Lee G, Yoon SJ, Shin BK, et al. UPLC-QTOFMS based metabolomics followed by stepwise partial least square-discriminant analysis (PLS-DA) explore the possible relation between the variations in secondary metabolites and the phylogenetic divergences of the genus Panax. J Chromatogr B Analyt Technol Biomed Life Sci 2016;1012–1013:61–8.
[114] Bruschi M, Granata S, Candiano G, Fabris A, Petretto A, Ghiggeri GM, et al. Proteomic analysis of urinary extracellular vesicles reveals a role for the complement system in medullary sponge kidney disease. Int J Mol Sci 2019;20.
[115] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J Chemom 2010;16:119–28.
[116] Wold S, Antti H, Lindgren F, Ohman J. Orthogonal signal correction of near-infrared spectra. Chemometr Intel Lab 1998;44:175–85.
[117] Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom 2006;20:341–51.
[118] Boccard J, Rutledge DN. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. Anal Chim Acta 2013;769:30–9.
[119] Duangkumpha K, Stoll T, Phetcharaburanin J, Yongvanit P, Thanan R, Techasen A, et al. Urine proteomics study reveals potential biomarkers for the differential diagnosis of cholangiocarcinoma and periductal fibrosis. PLoS ONE 2019;14:e0221024.
[120] Le Cao KA, Martin PG, Robert-Granie C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinf 2009;10:34.
[121] Jiang M, Wang C, Zhang Y, Feng Y, Wang Y, Zhu Y. Sparse partial-least-squares discriminant analysis for different geographical origins of Salvia miltiorrhiza by (1) H-NMR-based metabolomics. Phytochem Anal 2014;25:50–8.
[122] Cao KAL, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinf 2011;12:253.
[123] Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 2010;11:94.
[124] Grunwald NJ, Goss EM. Evolution and population genetics of exotic and re-emerging pathogens: novel tools and approaches. Annu Rev Phytopathol 2011;49:249–67.
[125] Jelonek J, Stefanowski J. Feature subset selection for classification of histological images. Artif Intell Med 1997;9:227–39.
[126] Mustaqeem A, Anwar SM, Majid M, Khan AR. editors. Wrapper method for feature selection to classify cardiac arrhythmia. Annual international conference of the IEEE Engineering in Medicine and Biology Society, 2017.
[127] Holland JH. Adaption in natural and artificial systems: an introductory analysis with applications to biology, control and AI. U Michigan Press; 1975.
[128] Mitchell M. An introduction to genetic algorithms. MIT Press; 1996.
[129] Zhu F, Han LY, Chen X, Lin HH, Ong S, Xie B, et al. Homology-free prediction of functional class of proteins and peptides by support vector machines. Curr Protein Pept Sci 2008;9:70–95.
[130] Akbari R, Ziarati K. A multilevel evolutionary algorithm for optimizing numerical functions. Int J Ind Eng Comput 2011;2:419–30.
[131] Zhu F, Han L, Zheng C, Xie B, Tammi MT, Yang S, et al. What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. J Pharmacol Exp Ther 2009;330:304–15.
[132] Chen JB, Chuang LY, Lin YD, Liou CW, Lin TK, Lee WC, et al. Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. Mitochondrial DNA 2014;25:231–7.
[133] Cohen WW, Greiner R, Schuurmans D, editors. Probabilistic hill-climbing. The workshop on computational learning theory and natural learning systems. 1994.
[134] Laskaris R. Artificial Intelligence: a modern approach. Library J 2015;140.
[135] Hernando L, Mendiburu A, Lozano JA, editors. Hill-Climbing algorithm: let's go for a walk before finding the optimum. Congress on evolutionary computation. 2018;1-7.
[136] Su SC, Lin CJ, Ting CK. An effective hybrid of hill climbing and genetic algorithm for 2D triangular protein structure prediction. Proteome Sci 2011;9 (1):1–8.
[137] Zhu F, Ma XH, Qin C, Tao L, Liu X, Shi Z, et al. Drug discovery prospect from untapped species: indications from approved natural product drugs. PLoS ONE 2012;7:e39782.
[138] Vergara JR, Estévez PAJNC. A review of feature selection methods based on mutual information. Neural Comput Appl 2014;24:175–86.
[139] Mao KZ. Orthogonal forward selection and backward elimination algorithms for feature subset selection. IEEE Trans Syst Man Cybern B Cybern 2004;34:629–34.
[140] Valsan S. Backward sequential feature elimination and joining algorithms in machine learning. San Jose State University; 2014.
[141] Theodoridis S, Koutroumbas K. Pattern recognition. 2nd ed. San Diego, USA: Elsevier Academic Press; 2003.
[142] Sharma A, Imoto S. Miyano SJIAToCB. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans Comput Biol Bioinform 2012;9:754–64.
[143] Ang JC, Mirzal A, Haron H, Hamed HN. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinform 2016;13:971–89.
[144] Figueroa A. Exploring effective features for recognizing the user intent behind web queries. Comput Ind 2015;68:162–9.
[145] Figueroa A, Neumann G. Category-specific models for ranking effective paraphrases in community question answering. Expert Syst Appl 2014;41:4730–42.
[146] Larraanaga P, Lozano JA. Estimation of distribution algorithms: a new tool for evolutionary computation. Kluwer Academic Publishers; 2001.
[147] Pelikan M, Goldberg DE, Lobo FG. A survey of optimization by building and using probabilistic models. Am Control Conf 2000;21(1):5–20.
[148] Pelikan M. Probabilistic model-building genetic algorithms. Berlin Heidelberg: Springer; 2005.
[149] Kim K, Shan Y, Nguyen XH, Mckay RI. Probabilistic model building in genetic programming: a critical review. Genet Program Evol M 2014;15:115–67.
[150] Bošković B, Brest JJASC. Genetic algorithm with advanced mechanisms applied to the protein structure prediction in a hydrophobic-polar model and cubic lattice. Appl Soft Comput 2016;45:61–70.
[151] Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved?. Curr Opin Struct Biol 2007;17:342–6.
[152] Su R, Liu X, Wei L, Zou Q. Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. Methods 2019;166:91–102.
[153] Khachaturyan AG, Semenovskaya SV, Vaĭnshteĭn BK. Statistical-thermodynamic approach to determination of structure amplitude phases. Soviet Phys Cryst 1979;24:905–16.
[154] Adabor ES, Acquaah-Mensah GKJ. Machine learning approaches to decipher hormone and HER2 receptor status phenotypes in breast cancer. Brief Bioinform 2017;20:504–14.
[155] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–17.
[156] Blaise BJ, Correia G, Tin A, Young JH, Vergnaud AC, Lewis M, et al. Power analysis and sample size determination in metabolic phenotyping. Anal Chem 2016;88:5179–88.
[157] Wan X, Liu J, Cheung WK, Tong T. Learning to improve medical decision making from imbalanced data without a priori cost. BMC Med Inf Decis Making 2014;14:111.
[158] Park J, Li K, Zhou H. K-fold subsampling based sequential backward feature elimination. In: International conference on pattern recognition applications and methods. p. 423–30.
[159] Wang Y, Makedon FS, Ford JC, Pearlman J. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 2005;21:1530–7.
[160] Yan R, Ma Z, Zhao Y, Kokogiannakis GJE. A decision tree based data-driven diagnostic strategy for air handling units. Energy Buildings 2016;133:37–45.
[161] Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, et al. Update of TTD: therapeutic target database. Nucleic Acids Res 2010;38:D787–91.
[162] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[163] Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 2012;67:93–104.
[164] Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.
[165] Zhou L, Wang Q, Yin P, Xing W, Wu Z, Chen S, et al. Serum metabolomics reveals the deregulation of fatty acids metabolism in hepatocellular carcinoma and chronic liver diseases. Anal Bioanal Chem 2012;403:203–13.
[166] Zeng W, Wang F, Ma Y, Liang X, Chen P. Dysfunctional mechanism of liver cancer mediated by transcription factor and non-coding RNA. Curr Bioinform 2019;14:100–7.
[167] Smola AJ, Scholkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222.
[168] Borgwardt KM, Ong CS, Schonauer S, Vishwanathan SV, Smola AJ, Kriegel HP. Protein function prediction via graph kernels. Bioinformatics 2005;21:47–56.
[169] Bottou L, Vapnik V. Local learning algorithms. Neural Comput 1992;4:888–900.
[170] Shao JL, Xu D, Tsai SN, Wang YF, Ngai SM. Computational identification of protein methylation sites through bi-profile bayes feature extraction. PLoS ONE 2009;4:e4920.
[171] Mladenic D, Grobelnik M. Feature selection on hierarchy of web documents. Decis Support Syst 2003;35:45–87.
[172] Chen J, Huang H, Tian S, Qu Y. Feature selection for text classification with naïve bayes. Expert Syst Appl 2009;36:5432–5.
[173] Mahadevan S, Shah SL, Marrie TJ, Slupsky CM. Analysis of metabolomic data using support vector machines. Anal Chem 2008;80:7562–70.
[174] Gaul DA, Mezencev R, Long TQ, Jones CM, Benigno BB, Gray A, et al. Highly-accurate metabolomic detection of early-stage ovarian cancer. Sci Rep 2015;5:16351.
[175] Hu X, Cammann H, Meyer HA, Miller K, Jung K, Stephan C. Artificial neural networks and prostate cancer-tools for diagnosis and management. Nat Rev Urol 2013;10:174–82.

[176] Sarve A, Sonawane SS, Varma MN. Ultrasound assisted biodiesel production from sesame (*Sesamum indicum L.*) oil using barium hydroxide as a heterogeneous catalyst: comparative assessment of prediction abilities between response surface methodology (RSM) and artificial neural network (ANN). Ultrason Sonochem 2015;26:218–28.
[177] Azadi S, Karimi-Jashni A. Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: a case study of fars province. Iran Waste Manag 2016;48:14–23.
[178] Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49:1225–31.
[179] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet 2019;51:12–8.
[180] Hong J, Luo Y, Mou M, Fu J, Zhang Y, Xue W, et al. Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. Brief Bioinform 2019. https://doi.org/10.1093/bib/bbz120.
[181] Hong J, Luo Y, Zhang Y, Ying J, Xue W, Xie T, et al. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. Brief Bioinform 2020;21:1437–47.
[182] Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer diagnosis using deep learning: a bibliographic review. Cancers (Basel) 2019;11.
[183] Kim H, Kim Y, Han B, Jang JY, Kim Y. Clinically applicable deep learning algorithm using quantitative proteomic data. J Proteome Res 2019;18:3195–202.
[184] Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the mouse cell atlas by microwell-seq. Cell 2018;173:1307.
[185] Agarwal R, Narayan J, Bhattacharyya A, Saraswat M, Tomar AK. Gene expression profiling, pathway analysis and subtype classification reveal molecular heterogeneity in hepatocellular carcinoma and suggest subtype specific therapeutic targets. Cancer Genet 2017;216–217:37–51.
[186] Liu W, Yang X, Wang N, Fan S, Zhu Y, Zheng X, et al. Multiple immunosuppressive effects of CpG-c41 on intracellular TLR-mediated inflammation. Mediators Inflamm 2017;2017:6541729.
[187] Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol 2010;28:83–9.
[188] Alonso A, Marsal S, Julia A. Analytical methods in untargeted metabolomics: state of the art in 2015. Front Bioeng Biotechnol 2015;3:23.
[189] Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. Comput Struct Biotechnol J 2013;4: e201301009.
[190] Platzer A. Visualization of SNPs with t-SNE. PLoS ONE 2013;8:e56883.
[191] Wang M, Kornblau SM, Coombes KR. Decomposing the apoptosis pathway into biologically interpretable principal components. Cancer Inform. 2018;17:1176935118771082.
[192] Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature 2015;526:131–5.
[193] Jin J, Son M, Kim H, Kim H, Kong SH, Kim HK, et al. Comparative proteomic analysis of human malignant ascitic fluids for the development of gastric cancer biomarkers. Clin Biochem 2018;56:55–61.
[194] Want EJ, Wilson ID, Gika H, Theodoridis G, Plumb RS, Shockcor J, et al. Global metabolic profiling procedures for urine using UPLC-MS. Nat Protoc 2010;5:1005–18.
[195] Nicholson JK, Connelly J, Lindon JC, Holmes E. Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 2002;1:153–61.
[196] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[197] Abdelmoula WM, Balluff B, Englert S, Dijkstra J, Reinders MJ, Walch A, et al. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. Proc Natl Acad Sci U S A 2016;113:12244–9.
[198] Beckonert O, Keun HC, Ebbels TM, Bundy J, Holmes E, Lindon JC, et al. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nat Protoc 2007;2:2692–703.
[199] Peng Y, Li X, Wu M, Yang J, Liu M, Zhang W, et al. New prognosis biomarkers identified by dynamic proteomic analysis of colorectal cancer. Mol BioSyst 2012;8:3077–88.
[200] Constantinou C, Chrysanthopoulos PK, Margarity M, Klapa MI. GC-MS metabolomic analysis reveals significant alterations in cerebellar metabolic physiology in a mouse model of adult onset hypothyroidism. J Proteome Res 2011;10:869–79.
[201] Kanonidis EI, Roy MM, Deighton RF, Le Bihan T. Protein co-expression analysis as a strategy to complement a standard quantitative proteomics approach: case of a glioblastoma multiforme study. PLoS ONE 2016;11:e0161828.
[202] Widlak P, Mrukwa G, Kalinowska M, Pietrowska M, Chekan M, Wierzgon J, et al. Detection of molecular signatures of oral squamous cell carcinoma and normal epithelium - application of a novel methodology for unsupervised segmentation of imaging mass spectrometry data. Proteomics 2016;16:1613–21.
[203] Kim S. Weighted K-means support vector machine for cancer prediction. Springerplus 2016;5:1162.
[204] Guo J, Jing R, Zhong JH, Dong X, Li YX, Liu YK, et al. Identification of CD14 as a potential biomarker of hepatocellular carcinoma using iTRAQ quantitative proteomics. Oncotarget 2017;8:62011–28.
[205] Kohonen T, Kaski S, Lagus K, Salojarvi J, Honkela J, Paatero V, et al. Self organization of a massive document collection. IEEE Trans Neural Netw 2000;11:574–85.
[206] Meinicke P, Lingner T, Kaever A, Feussner K, Gobel C, Feussner I, et al. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. Algorithms Mol Biol 2008;3:9.