


A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies

Jing Tang[†], Yunxia Wang[†], Jianbo Fu, Ying Zhou, Yongchao Luo, Ying Zhang, Bo Li, Qingxia Yang, Weiwei Xue, Yan Lou, Yunqing Qiu and Feng Zhu 

Corresponding authors: Feng Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China. Tel.: +86-571-88208444; E-mail: zhufeng@zju.edu.cn; Yunqing Qiu, The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang 310000, China. Tel.: +86-571-88236626; E-mail: qiuyq@zju.edu.cn

[†]These authors contributed equally to this work as co-first authors.

Abstract

Microbial community (MC) has great impact on mediating complex disease indications, biogeochemical cycling and agricultural productivities, which makes metaproteomics powerful technique for quantifying diverse and dynamic composition of proteins or peptides. The key role of biostatistical strategies in MC study is reported to be underestimated, especially the appropriate application of feature selection method (FSM) is largely ignored. Although extensive efforts have been devoted to assessing the performance of FSMs, previous studies focused only on their classification accuracy without considering their ability to correctly and comprehensively identify the spiked proteins. In this study, the performances of 14 FSMs were comprehensively assessed based on two key criteria (both sample classification and spiked protein discovery) using a variety of metaproteomics benchmarks. First, the classification accuracies of those 14 FSMs were evaluated. Then, their abilities in identifying the proteins of different spiked concentrations were assessed. Finally, seven FSMs (FC, LMEB, OPLS-DA, PLS-DA, SAM, SVM-RFE and T-Test) were identified as performing consistently superior or good under both criteria with the PLS-DA performing consistently superior. In summary, this study served as comprehensive analysis on the performances of current FSMs and could provide a valuable guideline for researchers in metaproteomics.

Key words: metaproteomics; feature selection method; classification accuracy; spiked proteins; microbiome

Jing Tang, Yunxia Wang, Jianbo Fu, Ying Zhou, Yongchao Luo, Ying Zhang, Bo Li and Qingxia Yang are the Master or PhD candidates of the College of Pharmaceutical Sciences at Zhejiang University, China and jointly cultivated by the School of Pharmaceutical Sciences at Chongqing University, China.

Weiwei Xue is an associate professor of the School of Pharmaceutical Sciences at Chongqing University, China. He is interested in the area of computational biology and molecular dynamics.

Yan Lou is an associate professor of the First Affiliated Hospital at Zhejiang University, China. She is interested in the area of clinical pharmacology, precision medicine and bioinformatics.

Yunqing Qiu is a professor of the First Affiliated Hospital at Zhejiang University, China. He is interested in the area of precision medicine, diagnosis and treatment of liver disease and system biology.

Feng Zhu is a professor of the College of Pharmaceutical Sciences at Zhejiang University, China. He got his PhD degree from the National University of Singapore, Singapore. His research group (<https://idrblab.org/>) has been working in the fields of bioinformatics, OMIC-based drug discovery, system biology and medicinal chemistry. All are welcome to visit his personal website at <https://idrblab.org/Peoples.php>.

Submitted: 27 March 2019; Received (in revised form): 14 April 2019

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

Microbial community (MC) has great impacts on mediating the global-scale biogeochemical cycling [1–3], revealing the pathogenesis of various diseases [4–7] and enhancing the agricultural productivities [8]. Investigation of the protein abundance in a given MC has enabled an unprecedented view of the adaptive response of microbes to the external stimuli or their interactions with other organism/host cells [9]. This makes metaproteomics powerful technique for quantifying diverse and dynamic composition of proteins or peptides [10–13]. Due to the enormous demands on discovering diagnostic/prognostic protein markers of high specificity and sensitivity [14], various feature selection methods (FSMs) have been constructed and applied to identify the relationship between microbe and host phenotype [15], facilitate the diagnosis of cognitive dysfunctions [16] and achieve the rapid and accurate identification of infectious strains that is essential for appropriate therapeutic management and timely intervention [17–20].

However, there are substantial challenges available in current metaproteomics researches, which include the heterogeneity among microbial samples [21–23], expanded search space [24–26], vast dynamic range of protein abundances [24, 27], great variation among sample preparations or experimental runs [28–30], complex processing steps [31, 32] and the absence of effective bioinformatics tool [31, 33]. Among these challenges, the key role of biostatistical strategy in microbiome study is found to be underestimated [34], especially the appropriate application of FSMs is largely ignored [34–36]. Particularly, the classification performance of the discovered features and the identification rate of the spiked proteins by several FSMs are found to be poor [37]. On one hand, the FSM should identify a minimal and optimal subset of features (e.g. the serum markers in predictive medicine) that are relevant to the development of prediction model with high accuracy [38], but its performances are frequently hampered, which decrease the quality of the feature discovery prior to experiment validation [39, 40]. On the other hand, an ideal FSM should select the complete list of discriminative features (e.g. the spiked proteins of known concentrations and spiked at different levels into background samples have been widely applied for analyzing the performances of various statistical methods in identifying the true difference in protein abundances) [14, 41, 42], but the sets of features identified by different FSMs are found sharing little in common [43–45]. Therefore, it is essential in current metaproteomics studies to simultaneously improve the performances of FSMs in both classification and spiked protein discovery [14].

Till now, no less than 14 FSMs have been developed and applied to analyze the mass spectrometry (MS)-based microbial proteomics data, which can be classified into the following two groups [46–48]: univariate filter and wrappers/embedded methods. The univariate filter FSMs (such as student's *t*-test and Wilcoxon rank-sum test) assess the relevance of features by looking only at the intrinsic properties of data but not for sample classification [49]. The wrappers/embedded FSMs select a list of features that provides the best classification for predefined sample groups. Since it remains unclear which FSM performs the best in recent metaproteomics [34], the comparisons among FSMs are reported [14]. Particularly, the abilities of six FSMs in clinical proteomics-based marker discovery are compared [14]; the capacities of three feature selection and three classification methods in enhancing information from overnight oximetry in the context of apnea diagnosis are assessed [50]. Some novel methods are compared with available ones on realistic microbial

data sets [51]. However, the performance assessments in these reported metaproteomics studies focus only on sample classification, and FSMs' performances on identifying spiked proteins are not fully considered. Moreover, no more than six FSMs have been assessed in previous publications. Therefore, it is critical to comprehensively assess the performances of all available FSMs for current metaproteomics by systematically considering both classification and spiked protein discovery.

In this study, the performances of 14 FSMs were comprehensively assessed from two perspectives (both sample classification and spiked protein discovery) using a variety of metaproteomics benchmark data sets. First, three metaproteomics benchmark data sets (each containing two distinct sample groups) were collected. Second, the classification performances of those 14 FSMs were evaluated using these benchmarks. Third, another 3 benchmark data sets with the spiked proteins of 5 different concentrations were collected, and 10 subsequent data sets of different concentration combinations were generated. Finally, the abilities of FSMs in identifying the proteins of different spiked concentrations were assessed. All in all, this study served as a comprehensive analysis on the performances of current popular FSMs and could provide a valuable guideline for the researchers in the field of metaproteomics.

Materials and methods

Collection of the metaproteomics benchmark data sets

To enable a comprehensive assessment on FSMs, a variety of metaproteomics benchmark data sets were collected from the PRIDE database [52] by searching the keywords 'Metaproteomic', 'Microbiota' and 'Microbiome', which resulted in 106 records relevant to metaproteomics. The corresponding literatures of the resulting records were then comprehensively reviewed. By considering several additional criteria [the label-free quantification (LFQ), the availability of raw intensity data file and the protein database or library to search against, the well-defined parameters such as isolation scheme and range of retention time, the clear description on distinct sample groups and so on], 10 metaproteomics benchmark data sets were then identified. Literature reviews discovered one additional metaproteomics benchmark data set (the CPTAC dataset [53]; three technical replicates of UPS1 proteins spiked into a yeast proteome digesting with a variety of concentrations—0.25, 0.74, 2.2, 6.7 and 20 fmol/ μ l). As shown in Table 1, 11 benchmark data sets were collected in total, and the number of samples and a brief description on each data set were provided. In order to effectively assess the classification performance of each FSM, only the benchmarks with >10 samples in either sample groups [54] were included in the analyses of this study, which resulted in three datasets: PXD006224 [55] (60 'metabolic phase' and 24 'equilibrium phase' fecal samples); PXD002882 [56] (21 Crohn's disease patients and 10 control healthy subjects); PXD006129 [57] (14 western-style diet and 14 chow-fed mice). Moreover, to accurately evaluate the identification performance of spiked proteins, only the benchmark data sets spiked with multiple proteins were analyzed, which brought about another three sets of metaproteomics data (PXD002099 [58], PXD001819 [41] and CPTAC-ST6 [53], as shown in Table 1). There were eight data sets included in the CPTAC study [53], and only the fifth and sixth data sets contained the spiked proteins. The fifth data set only spiked by one protein of single concentration, which was not a control-case study and not suitable for the analysis. Thus, only

Table 1. Description and statistics of the benchmark metaproteomics data sets collected for the analysis of this study

Data sets	Dataset ID	Dataset description	Spiked proteins	Number of proteins	Quantification software tool
<i>J Proteome Res.</i> 14:4118–26, 2015	PXD002099	The UPS1 mixtures of different concentrations (2, 4, 10, 25 and 50 fmol/μL) were spiked into yeast proteome [three runs for each concentration]	48 spiked proteins	1442	Progenesis
<i>Data Brief.</i> 6:286–94, 2014	PXD001819	The Sigma UPS1 of different concentrations (0.5, 5, 12.5, 25 and 50 fmol/μg) were spiked into a background of yeast lysate [three runs for each concentration]	48 spiked proteins	868	MaxQuant
<i>J Proteome Res.</i> 9:761–76, 2010	CPTAC-ST6	The UPS1 proteins of different concentrations (0.25, 0.74, 2.2, 6.7 and 20 fmol/μL) were spiked into a yeast proteome [three runs for each concentration]	48 spiked proteins	1570	MaxQuant
<i>Microbiome.</i> 5:144, 2017	PXD006224	60 metabolic phase fecal samples; 24 equilibrium phase fecal samples	No spiked proteins	9761	MaxQuant
<i>Nat Commun.</i> 7:13419, 2016	PXD002882	21 Crohn's disease patients; 10 control subjects	No spiked proteins	4169	MaxQuant
<i>Cell Host Microbe.</i> 23:27–40, 2018	PXD006129	14 western-style diet mice; 14 chow-fed mice	No spiked proteins	3243	MaxQuant
<i>Front Microbiol.</i> 8:1605, 2017	PXD006070	nine corn silage-based samples; nine grass silage-based samples	No spiked proteins	8163	MaxQuant
<i>Genome Med.</i> 8:44, 2016	PXD003028	eight people before breakfast; eight people after breakfast	No spiked proteins	6431	MaxQuant
<i>Mol Cell Proteomics.</i> 13:2277–87, 2014	PXD000987	four transverse colon samples; four descending colon samples	No spiked proteins	2817	MaxQuant
<i>Front Microbiol.</i> 8:1215, 2017	PXD005929	three surfaces exposed; three whole cell extracts	No spiked proteins	1570	MaxQuant
<i>J Biol Chem.</i> 292:17337–50, 2017	PXD006810	three NleB1-infected cells; three wild-type cells	No spiked proteins	1195	MaxQuant

the sixth dataset (CPTAC-ST6, with five different concentrations) was selected as one of the three benchmarks analyzed here. All in all, each of these three data sets was spiked with UPS1 proteins of five different concentrations, and any two concentrations could be combined to form a pair with two distinct groups, which resulted in 10 different pairs of data sets. Taking the PXD002099 as an example, 10 pairs of data sets with 10 different ratios of concentration (2 versus 4, 2 versus 10, 2 versus 25, 2 versus 50, 4 versus 10, 4 versus 25, 4 versus 50, 10 versus 25, 10 versus 50 and 25 versus 50 fmol/μL) were generated by randomly combining any two concentrations in PXD002099.

Preprocessing of the collected metaproteomics data

The raw LFQ intensities of proteins were downloaded from ProteomeXchange repository [59]. The metaproteomics data were reported to be characterized by sparsity, which could be represented by a substantial amount of missing values (~40%) and affected up to 80% of protein features [60]. To cope with these problems, the metaproteomics data sets collected in this study

were preprocessed using the following four steps: (i) the LFQ intensities were extracted from the downloadable raw data files; (ii) the MaxQuant [61] was then applied to generate protein group data files by setting the false discovery rate to 0.01 and requiring each protein group to have at least one unique or razor peptide [62]; (iii) the K-nearest neighbor (KNN) algorithm was applied to impute the missing intensities [63]; and (4) all LFQ intensities were finally normalized using the variance stabilization normalization (VSN) method [64]. In particular, the missing values occurred frequently in proteomics data set and critically affected the downstream analyses [65, 66]; it was thus a common practice to impute these missing values. The KNN imputation aimed at identifying K proteins that were similar to the proteins with missing values, where the similarity was estimated by Euclidean distance, and the missing values were imputed with the values of weighted average from the neighboring proteins [67]. KNN was reported to outperform other imputation methods (such as Bayesian principal component analysis and local least squares) in current proteomics study [67]; it was thus adopted to impute the missing values in this study. Moreover, the normalization was essential for metaproteomics analysis that aimed at

reducing systematic bias or technical variations for improving the comparability of data and the reliability for downstream analyses [68–70]. Among all normalizations, the VSN was one of the most popular approaches [65, 68, 71]. VSN could not only reduce the unwanted variations among technical replicates in all examined data sets [68] but also performed consistently well in the differential expression analysis [65]; it was therefore adopted to normalize the benchmark datasets in this study.

The FSMs assessed in this study

In total, 14 FSMs popular in current metaproteomics studies were assessed in this study, which included (i) Chi-square (CHIS): judging the independence of two events and being erratic for very small expected count [72]; (ii) empirical Bayesian analysis of microarray (EBAM): identifying differentially expressed genes and being used in many multiple testing situations [73]; (iii) entropy-based filters (ENTROPY): filter-based feature ranking techniques including information gain, gain ratio and symmetrical uncertainty [74]; (iv) fold change (FC): generating more reproducible features than the ordinary and modified t-statistics [75]; (v) linear models and empirical Bayes (LMEB): assessing the differential intensities by measuring features based on t-statistics and fold changes simultaneously [76]; (vi) orthogonal partial least squares discriminant analysis (OPLS-DA): an upgraded version of PLS-DA method to discriminate two or more groups using the multivariate data [77]; (vii) partial least squares discriminant analysis (PLS-DA): the chemometrics technique used for classification purposes by trying to find a proper compromise between describing the data set and predicting the response [78]; (viii) random forest (RF): an ensemble, supervised machine-learning algorithm for pattern recognition in OMIC data sets [79]; (ix) random forest—recursive feature elimination (RF-RFE): recursive backward feature elimination procedure [80]; (x) significance analysis for microarrays (SAM): permutation-based (non-parametric) hypothesis testing method for the identification of quantities that differ greatly between two measurement sets [81, 82]; (xi) sparse partial least squares discriminant analysis (SPLS-DA): a direct application of sparse partial least squares with competitive computational efficiencies and interpretability of results via valuable graphical outputs [83]; (xii) support vector machine—recursive features elimination (SVM-RFE): wrapper FSM generating the ranking of features via backward feature eliminations [84]; (xiii) student's t-test (T-Test): one of the most prevalent tests used in medical field and the most powerful unbiased test under normal curve theory [85]; (xiv) Wilcoxon rank-sum test (Wilcox): a non-parametric method under extremely skewed distribution [85]. Detailed information of each FSM was provided in the [Supplementary Methods](#).

Assessing FSMs' performance by classification accuracy and identified spiked proteins

Classification accuracy was used to judge the reliability of the selected biomarker candidates [14], which was applied in this study to assess FSMs' performances. First, the discriminative proteins were identified and ranked using the FSMs. Then, the top-ranked proteins (top 20, top 50, top 100, top 150, top 200, top 250, top 300, top 350, top 400 and top 450) were identified. Third, SVM [86] was applied to assess the performances of FSMs based on 10 different sets of top-ranked proteins using 5-fold cross validation. Considering the influence of parameters used in the

machine-learning algorithms [87], all SVM models constructed based on the features derived by the FSMs underwent the process of parameter optimization in this study. All calculations were conducted using R (<http://www.r-project.org>) version 3.5.3 running on Linux v6.5 operating system of 128GB RAM and CPU E7-4820 × 32 cores.

An ideal FSM should screen a complete list of differential features that are related to the spiked proteins [88, 89]. In this study, the performances of 14 FSMs were thus evaluated by measuring each algorithm's capacity of constructing an optimal feature set, which would only contain those features related to spiked peptides (true positives). To accomplish this, three benchmark data sets (PXD002099 [58], PXD001819 [41] and CPTAC-ST6 [53], shown in [Table 1](#)) with spiked proteins were first collected. Then, each FSM was used to three benchmarks for identifying different sets of features, which contained spiked proteins (true positives) and non-spiked proteins (false positives). Finally, the total number of the spiked proteins identified by each FSM was used to assess the performance of FSMs [14].

The relationship among FSMs' performances identified by hierarchical clustering

The classification accuracy and the number of identified spiked proteins of the FSMs were calculated to assess their performances. The hierarchical clustering of the FSMs based on the above two metrics was applied to identify the relationship among the performances of various FSMs. On one hand, the accuracy values of a specific FSM among top N differentially expressed proteins (20, 50, 100, 150, 200, 250, 300, 350, 400 and 450) were used to generate a 10-dimensional vector. On the other hand, the number of spiked proteins identified by a specific FSM among 10 pairs of data sets with 10 different ratios of concentration could also be used to construct a 10-dimensional vector. Then, the hierarchical clustering was applied to explore the relationship among 14 vectors (corresponding to 14 FSMs). To measure the distance between any two vectors, the Euclidean distance was adopted, and the clustering method was the Ward's minimum variance [90], which could reduce the total within-cluster variance to the maximum extents. In this work, the Ward's minimum variance module in R package was used [91].

Results and discussion

Data preprocessing for removing unwanted variations

Unwanted experimental/biological variations may hamper the identification of differentially expressed proteins and affect the effectiveness of metaproteomics analysis [69, 92]. Thus, the VSN was applied to normalize the six benchmark data sets to remove specific types of unwanted variations, which was reported to be the one that can well reduce variation among studied samples [68]. The relative log abundances (RLA) plot and the pooled median absolute deviation (PMAD) were commonly accepted and widely used measures to assess the performance of normalization [93]. Compared with the RLA plot of unnormalized data ([Figure 1A](#)), the plots after VSN normalization ([Figure 1B](#)) gave a median closer to zero and lower variations around the median ([Supplementary Figures S1–S3](#)). Moreover, all benchmark data sets gave PMAD values less than 0.14 after VSN ([Figure 1C](#) and [Supplementary Figure S4](#)), which indicated a superior normalization [94]. All in all, the VSN performed very well on removing the unwanted variation as indicated by both RLA plot and PMAD values.

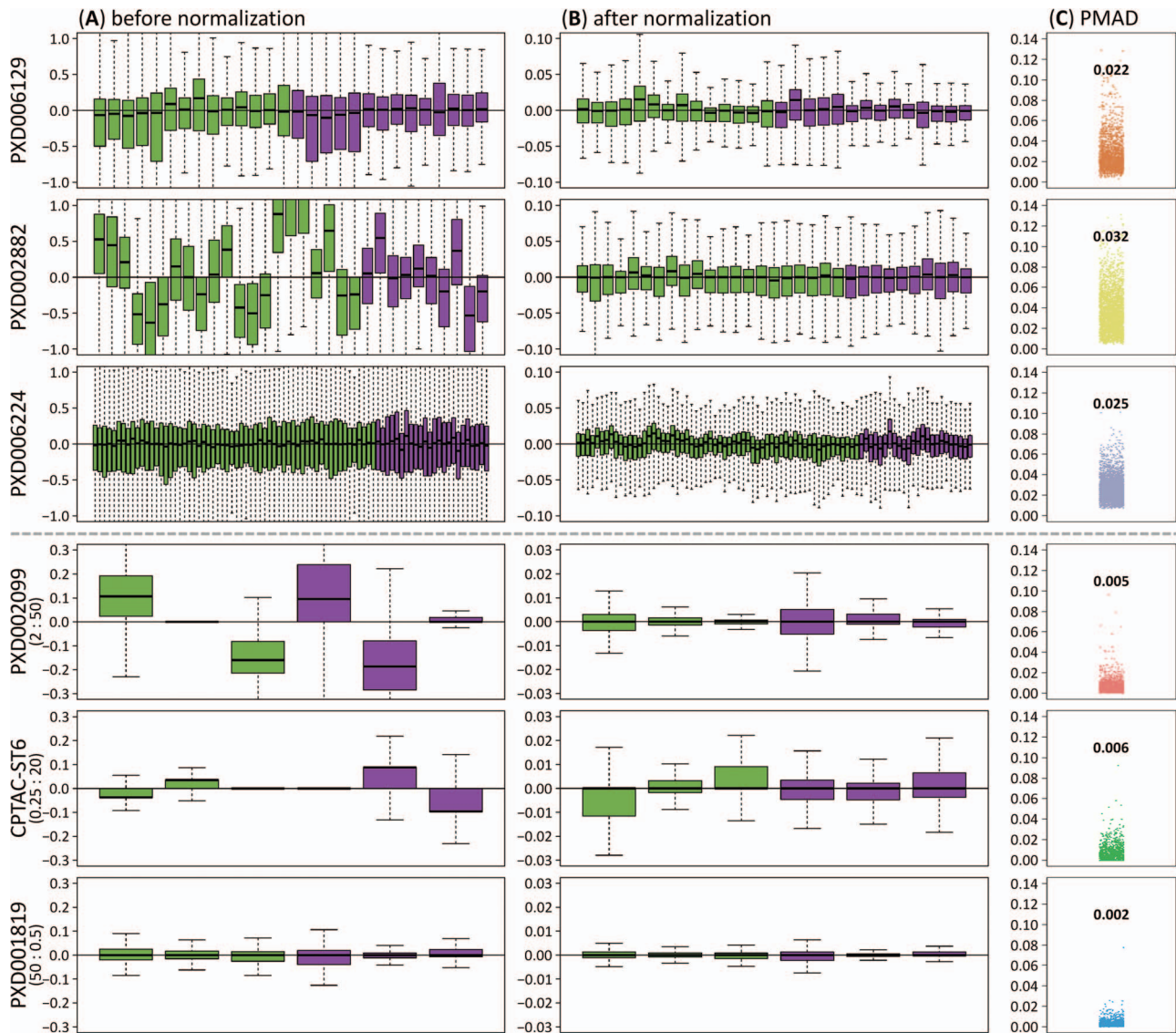


Figure 1. The RLA plots before and after the VSN normalization and the PMAD for six benchmark data sets. (A) The RLA plots for unnormalized intensities; (B) the RLA plots for normalized intensities; (C) the distributions of PMAD values. Purple: control group; Green: case group.

The performance of FSMs assessed by classification accuracies

An ideal FSM would be capable of identifying the optimal set of features with satisfactory classification accuracy (ACC). In this study, the ACCs of all 14 FSMs were assessed based on 5-fold cross validations [87, 95]. Table 2 provided the ACCs of all FSMs trained by 10 different sets of top-ranked proteins (top 20, top 50, top 100, top 150, top 200, top 250, top 300, top 350, top 400 and top 450) using 5-fold cross validation based on SVM classification. As shown, there were great variations among the ACCs of FSMs across three benchmark data sets. On one hand, the ACCs of a specific FSM across different data sets varied. Taking the CHIS (top 50) as an example, the ACCs of three benchmarks ranged from 0.677 (PXD002882) to 0.821 (PXD006224). On the other hand, the ACCs of various FSMs on a particular data set differed significantly. Taking the PXD006224 (top 50) as an example, the ACCs of different FSMs varied from 0.714 (Wilcox) to 0.929 (SPLS-DA, LMEB and EBAM). Because of this significant variation, it is of great interest

to identify FSMs consistently well-performing across multiple benchmarks and based on different sets of top-ranked protein features.

Identification of the FSMs with consistently high classification accuracies

The ACCs of a specific FSM among 10 different sets of top-ranked proteins were calculated to construct a 10-dimensional vector. Since the total number of features identified by RF-RFE was less than 10, there were 13 vectors corresponding to the remaining 13 FSMs. The hierarchic clustering of these 13 vectors resulted in Figure 2. As shown, 13 FSMs were divided by the dendrogram on the left side of each figure, which could be further grouped into three areas: top, middle and bottom colored in green, blue and gray, respectively. Clearly, three methods (SVM-RFE, SPLS-DA and PLS-DA) were consistently grouped to the top area across three benchmarks, while one method (Wilcox) always stayed in the bottom area. Thus, all FSMs could be further categorized to

Table 2. Assessing the performances of different FSMs based on their ACC value for three metaproteomics benchmark data sets (PXD006129, PXD002882 and PXD006224). ACC was calculated based on 5-fold cross validation, which was defined as (true positive + true negative)/(true positive + false positive + true negative + false negative)

FSM	PRIDE ID	The ACC values of different features size across three benchmark data sets									
		20	50	100	150	200	250	300	350	400	450
CHIS	PXD006129	0.571	0.679	0.536	0.714	0.679	0.536	0.679	0.786	0.857	0.821
	PXD002882	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.774	0.871
	PXD006224	0.786	0.821	0.845	0.869	0.857	0.881	0.881	0.881	0.929	0.929
EBAM	PXD006129	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
	PXD002882	0.677	0.677	0.677	0.677	0.710	0.742	0.774	0.839	0.903	0.903
	PXD006224	0.905	0.929	0.917	0.905	0.940	0.952	0.952	0.964	0.964	0.976
ENTROPY	PXD006129	0.750	0.643	0.571	0.643	0.786	0.679	0.643	0.679	0.786	0.857
	PXD002882	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.839	0.871
	PXD006224	0.786	0.810	0.869	0.857	0.881	0.869	0.881	0.917	0.929	0.929
FC	PXD006129	0.964	0.893	0.893	0.929	0.893	0.857	0.893	0.929	0.857	0.786
	PXD002882	0.677	0.677	0.742	0.710	0.774	0.774	0.774	0.839	0.871	0.806
	PXD006224	0.810	0.821	0.857	0.845	0.845	0.857	0.881	0.845	0.881	0.833
LMEB	PXD006129	0.929	0.893	0.964	0.964	0.893	0.929	0.929	0.786	0.786	0.750
	PXD002882	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.935	0.935
	PXD006224	0.917	0.929	0.929	0.929	0.929	0.940	0.940	0.940	0.940	0.976
OPLS-DA	PXD006129	0.893	0.821	0.893	0.821	0.821	0.821	0.786	0.750	0.786	0.750
	PXD002882	0.968	0.935	0.935	0.935	0.935	0.968	0.968	0.935	0.935	0.968
	PXD006224	0.917	0.905	0.929	0.929	0.940	0.940	0.940	0.952	0.952	0.940
PLS-DA	PXD006129	0.929	0.893	0.964	0.893	0.929	0.964	0.929	0.964	0.893	0.929
	PXD002882	0.935	0.968	0.935	0.968	0.968	0.968	0.968	0.968	0.968	0.968
	PXD006224	0.929	0.917	0.929	0.917	0.940	0.917	0.940	0.976	0.952	0.976
RF	PXD006129	0.893	0.929	0.964	0.893	0.929	0.857	0.964	0.964	0.857	0.821
	PXD002882	0.677	0.677	0.710	0.710	0.677	0.742	0.806	0.839	0.903	0.935
	PXD006224	0.774	0.786	0.786	0.821	0.857	0.881	0.893	0.905	0.940	0.952
SAM	PXD006129	0.929	0.929	0.929	0.893	0.929	0.821	0.964	0.821	0.857	0.857
	PXD002882	0.710	0.710	0.710	0.710	0.806	0.839	0.839	0.903	0.903	0.903
	PXD006224	0.821	0.833	0.845	0.857	0.869	0.881	0.869	0.869	0.869	0.857
SPLS-DA	PXD006129	0.893	0.964	0.893	0.929	0.964	0.893	0.929	0.964	0.821	0.929
	PXD002882	0.935	0.935	0.968	0.968	1.000	0.968	0.968	0.968	0.968	0.968
	PXD006224	0.917	0.929	0.929	0.917	0.917	0.940	0.917	0.952	0.952	0.964
SVM-RFE	PXD006129	1.000	1.000	0.964	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	PXD002882	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	PXD006224	0.929	0.905	0.917	0.929	0.917	0.952	0.964	1.000	0.988	1.000
T-Test	PXD006129	0.893	0.857	0.929	0.929	0.929	0.929	0.893	0.929	0.893	0.929
	PXD002882	0.935	0.935	0.935	0.968	0.968	0.968	0.968	0.968	0.935	0.935
	PXD006224	0.857	0.893	0.905	0.893	0.893	0.881	0.905	0.893	0.917	0.893
Wilcox	PXD006129	0.643	0.643	0.786	0.714	0.714	0.607	0.714	0.750	0.893	0.893
	PXD002882	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.742
	PXD006224	0.714	0.714	0.714	0.714	0.714	0.714	0.714	0.726	0.738	0.762

four classes by comprehensively considering their ACCs (Table 2 and Figure 2) across three benchmarks. As shown in Figure 3, the FSMs of orange boxes in class C-A (SVM-RFE, SPLS-DA and PLS-DA) provided the best classification accuracies among FSMs, which made this class of Superior performance. This result was partially consistent with previous publications that (i) the predictive performance of SVM-RFE was reported to be the best among other FSMs [34, 96] and (ii) PLS-DA was discovered as robust and well-performing method in proteomics for various sample size of biological samples [14]. Moreover, the remaining 10 methods could be further divided into C-B1 (including 6 FSMs occasionally grouped to the top area but absent in the bottom of Figure 2 and 1 FSM staying in the middle area of Figure 2, yellow boxes with Good performance), C-B2 (including 2 FSMs occasionally grouped to the bottom area of Figure 2, blue boxes with Fair performance) and C-C (including 1 FSM consistently performing the worst across all benchmarks in Figure 2, grey boxes with Poor performance). Although the FSMs

in C-B1 slightly underperformed compared with that in C-A, they resulted in very good ACCs across 10 different sets of top-ranked protein features. However, the remaining three FSMs in C-B2 and C-C were not as well as that in C-A and C-B1, especially Wilcox method. Moreover, the variation in ACCs across benchmarks may be attributed to existing individual statistical biases [97].

The performance of FSMs evaluated by the number of identified spiked proteins

Another important measure employed to characterize the performance of FSMs was the extent to which they successfully identified the whole set of spiked proteins [14]. In this study, the total number of spiked proteins identified by each FSM was therefore calculated based on differential expression analyses [14]. Figure 4 illustrated the distribution of the number of spiked proteins identified by each FSM. As shown, for three benchmark

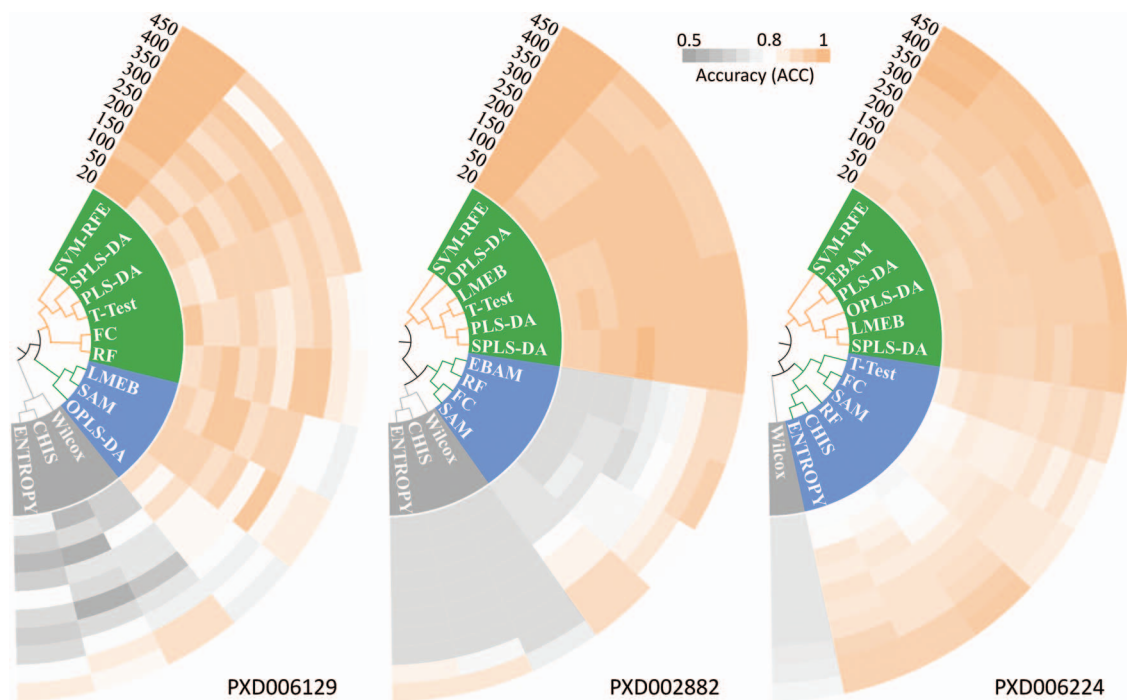


Figure 2. Clustering analysis of the studied FSMs using their classification accuracies (ACCs) across 10 different sets of top-ranked protein features based on 3 benchmark data sets (PXD006129, PXD002882 and PXD006224). The black-colored numbers indicated the number of protein features identified by FSMs (leaves of the hierarchical trees). Each cell in heat map represents the ACC values. The cell of the highest value was set as exact orange with the lower ones gradually fading towards gray (the lowest value).

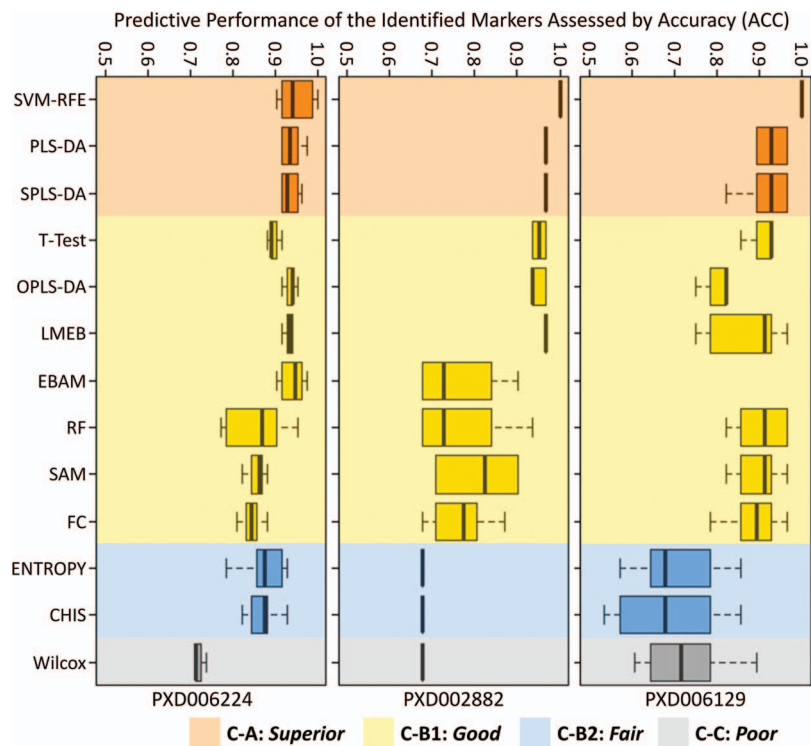


Figure 3. The classes of the studied FSMs defined in this study based on the top, middle and bottom areas in Figure 2. The FSMs in class A provided the best classification accuracies, which made the class A (C-A, orange boxes) with Superior performance. The FSMs in class C-B1 were occasionally classified to the top area but absent in the bottom of Figure 2 (yellow boxes with Good performance). The FSMs in C-B2 were occasionally classified to the bottom area of Figure 2 (blue boxes with Fair performance). The FSMs in class C-C performed consistently the worst across three benchmarks in Figure 2 (grey boxes with Poor performance).

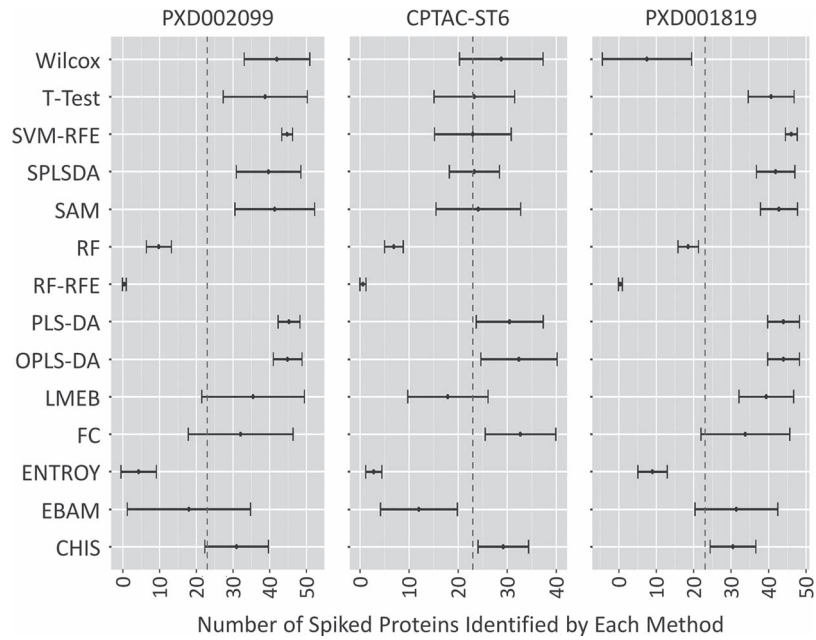


Figure 4. The distributions of the numbers of the spiked proteins identified by FSMs based on benchmark data sets (PXD002099, CPTAC-ST6 and PXD001819). The bar plots consisted of 10 pairs of data sets with 10 different ratios of concentration.

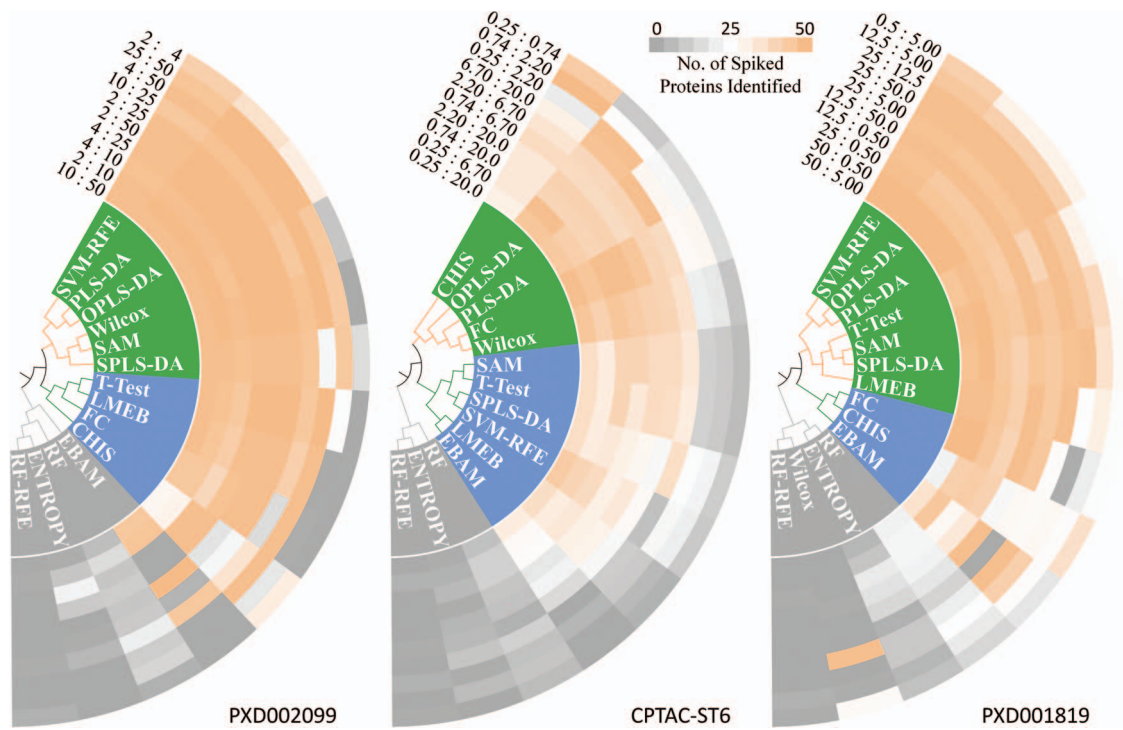


Figure 5. Clustering analysis of the studied FSMs using their numbers of identified spiked proteins across 10 different ratios of spiked protein concentrations based on 3 benchmark data sets (PXD002099, CPTAC-ST6 and PXD001819). The black-colored ratios indicated different ratios of spiked protein concentrations, and the FSMs were the leaves of the hierarchical trees. Each cell in the heat map represents the number of identified spiked proteins. The cell of the highest number was defined as exact orange with the lower ones gradually fading towards gray (the exact gray indicated zero).

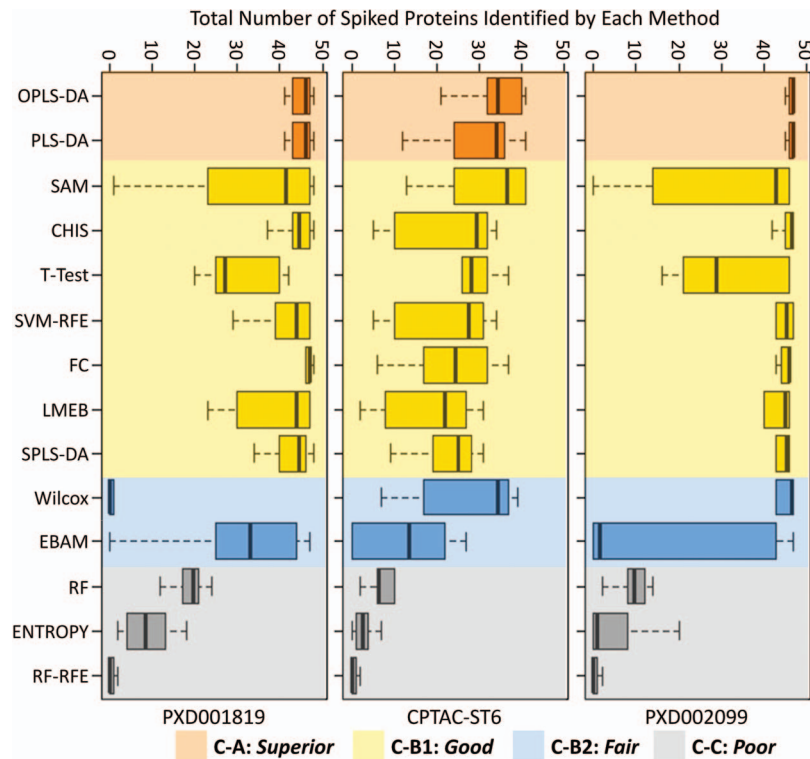


Figure 6. The classes of the studied FSMs defined in this study based on the top, middle and bottom areas in Figure 5. The FSMs in class A provided the highest number of identified spiked proteins, which made the class A (C-A, orange boxes) with Superior performance. FSMs in C-B1 were occasionally grouped to the top area but absent in the bottom of Figure 5 (yellow boxes with Good performance). The FSMs in C-B2 were occasionally classified to the bottom area of Figure 5 (blue boxes with Fair performance). FSMs in class C-C performed consistently the worst across three benchmarks in Figure 5 (grey boxes with Poor performance).

data sets, the distribution of the number of identified spiked proteins differed among FSMs. For PXD002099, the mean values of the numbers of spiked proteins identified by 14 FSMs were in the range from 0.4 (RF-RFE) to 45 (PLS-DA). Particularly, the mean values of 10 FSMs (CHIS, FC, LMEB, OPLS-DA, PLS-DA, SAM, SPLS-DA, SVM-RFE, T-test and Wilcox) were >24 , which denoted a high identification rate ($>50\%$) of spiked proteins by these FSMs. Meanwhile, the mean values of the remaining four FSMs (EBAM, ENTROPY, RF-RFE and RF) were <24 , and three FSMs (ENTROPY, RF-RFE and RF) were even smaller than 10, which indicated relatively low identification rate. For CPTAC-ST6, the mean values of the numbers of spiked proteins identified by the FSMs ranged from 0.6 (RF-RFE) to 32.7 (FC). Particularly, the mean values of nine FSMs (CHIS, FC, OPLS-DA, PLS-DA, SAM, SPLS-DA, SVM-RFE, T-test and Wilcox) were >24 , which were roughly consistent with the results of PXD002099. In the meantime, the mean values of the remaining five FSMs (EBAM, ENTROPY, LMEB, RF-RFE and RF) were <24 , and three FSMs (ENTROPY, RF-RFE and RF) were even smaller than 10, which were also relatively consistent with the results of PXD002099. For PXD001819, the mean values of total numbers of spiked proteins identified by all 14 FSMs ranged from 0.4 (RF-RFE) to 46.1 (SVM-RFE). Particularly, the mean values of 10 FSMs (CHIS, FC, LMEB, OPLS-DA, PLS-DA, SAM, SPLS-DA, SVM-RFE, T-test and Wilcox) were >24 , which was similar to that of PXD002099. Meanwhile, the mean value of the remaining four FSMs (ENTROPY, RF-RFE, RF and Wilcox) were <24 , and three FSMs (ENTROPY, RF-RFE and Wilcox) were even smaller than 10.

Discovery of the FSMs capable of consistently identifying high number of spiked proteins

The numbers of the spiked proteins identified based on 10 pairs of data sets with different concentration ratios by each FSM were used to construct a 10-dimensional vector. The resulting 14 vectors were then hierarchically clustered. As shown in Figure 5, 14 FSMs were divided by the corresponding dendrogram on the left side of each subfigure into three areas: top, middle and bottom colored in green, blue and gray, respectively. Clearly, two methods (OPLS-DA and PLS-DA) were consistently grouped into the top area across three benchmarks, while three methods (ENTROPY, RF and RF-RFE) always stayed in the bottom area. Thus, all FSMs could be further categorized into four classes by comprehensively considering their numbers of the identified spiked proteins (Figures 4 and 5) across three benchmark data sets. As illustrated in Figure 6, the FSMs of orange boxes in the class C-A (OPLS-DA and PLS-DA) performed consistently the best among all FSMs, which made this class with Superior performance. Moreover, the remaining 12 methods could be further divided into C-B1 (including seven FSMs occasionally grouped to the top but absent in the bottom of Figure 5, yellow boxes with Good performance), C-B2 (including two FSMs occasionally grouped to the bottom of Figure 5, blue boxes with Fair performance) and C-C (including three FSM consistently performing the worst across three benchmarks in Figure 5, grey boxes of Poor performance). Although the FSMs in C-B1 were slightly underperformed compared with that in the C-A, they identified

high number of spiked proteins across 10 pairs of data sets with different concentration ratios.

Generally, the protocol of processing metaproteomics data was organized into five sequential procedures [31]: (i) sample-specific database construction, (ii) protein identification and quantification, (iii) data preprocessing, (iv) statistical analyses and (v) protein taxonomy/function analysis. Particularly, the database construction/selection aimed at generating reduced, sample-specific protein database from original large databases [31]. MS data from microbe needed be searched against the constructed database based on various search engines for the peptide and protein identification. Meanwhile, various quantitative techniques could be used to measure the expression level of proteins, and protein abundance estimation counted the number of the identified protein [31]. Then, the data preprocessing consisted of data transformation, normalization and missing value imputation [98], which were frequently performed before the statistical analysis aiming at identifying the protein markers. Finally, function analyses aimed at investigating the enrichment of predefined groups with functionally related proteins [99].

Conclusions

Collective consideration of both classification accuracy and the identification of spiked proteins resulted in a comprehensive assessment of FSMs' performance. Based on the data of Figures 3 and 6, seven FSMs (FC, LMEB, OPLS-DA, PLS-DA, SAM, SVM-RFE and T-Test) performed consistently Superior or Good under both criteria, with the PLS-DA performing consistently Superior. Meanwhile, four FSMs (CHIS, EBAM, RF and SPLS-DA) were found to perform inconsistently under two criteria: (i) three FSMs (EBAM, RF and SPLS-DA) performed Superior or Good in their classification accuracy but performed Fair or Poor in the identification of spiked proteins; (ii) one FSM (CHIS) performed Good in identifying spiked proteins but with Fair classification accuracy. Moreover, two methods (ENTROPY and Wilcox) performed consistently Fair or Poor under both criteria. All in all, this study highlighted the importance of choosing appropriate FSMs in the biomarker discovery of metaproteomics study and identified several FSMs performing consistently well under two key criteria based on a variety of benchmark data sets.

Key Points

- The performances of feature selection methods (FSMs) in current metaproteomics studies were comprehensively assessed.
- The assessment was conducted based on two key criteria (sample classification and spiked protein discovery) using a variety of metaproteomics benchmarks.
- Seven FSMs were identified as performing consistently superior or good under both criteria, with the PLS-DA performing consistently superior.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

Funded by the National Key Research and Development Program of China (2018YFC0910500), the National Natural Science Foundation of China (81872798), Innovation Project on Industrial Generic Key Technologies of Chongqing (cstc2015zdcy-ztxx120003) and the Fundamental Research Funds for Central Universities (2018QNA7023, 10611CDJXZ238826, 2018CDQYSG0007, CDJZR14468801).

References

1. Arora-Williams K, Olesen SW, Scandella BP, et al. Dynamics of microbial populations mediating biogeochemical cycling in a freshwater lake. *Microbiome* 2018;6:165.
2. Roux S, Brum JR, Dutilh BE, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 2016;537:689–93.
3. Cui X, Yang Q, Li B, et al. Assessing the effectiveness of direct data merging strategy in long-term and large-scale pharmacometabonomics. *Front Pharmacol* 2019;10:127.
4. Duerkop BA, Kleiner M, Paez-Espino D, et al. Murine colitis reveals a disease-associated bacteriophage community. *Nat Microbiol* 2018;3:1023–31.
5. Wang L, Ping PY, Kuang LN, et al. A novel approach based on bipartite network to predict human microbe–disease associations. *Curr Bioinform* 2018;13:141–8.
6. Li YH, Li XX, Hong JJ, et al. Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief Bioinform* 2019, doi: [10.1093/bib/bby130](https://doi.org/10.1093/bib/bby130).
7. Yang H, Qin C, Li YH, et al. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;44:D1069–74.
8. Xu L, Naylor D, Dong Z, et al. Drought delays development of the sorghum root microbiome and enriches for monoderm bacteria. *Proc Natl Acad Sci U S A* 2018;115:E4284–93.
9. Broberg M, Doonan J, Mundt F, et al. Integrated multi-omic analysis of host-microbiota interactions in acute oak decline. *Microbiome* 2018;6:21.
10. Galand PE, Pereira O, Hochart C, et al. A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *ISME J* 2018;12:2470–8.
11. Li S, Li J, Ning L, 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.
12. Tang J, Fu J, Wang Y, et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform* 2019, doi: [10.1093/bib/bby127](https://doi.org/10.1093/bib/bby127).
13. Han Z, Xue W, Tao L, et al. Genome-wide identification and analysis of the eQTL lncRNAs in multiple sclerosis based on RNA-seq data. *Brief Bioinform* 2019, doi: [10.1093/bib/bbz036](https://doi.org/10.1093/bib/bbz036).
14. Christin C, Hoefsloot HC, Smilde AK, et al. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol Cell Proteomics* 2013;12:263–76.
15. Faith JJ, Ahern PP, Ridaura VK, et al. Identifying gut microbe–host phenotype relationships using combinatorial communities in gnotobiotic mice. *Sci Transl Med* 2014;6:220ra11.
16. Agarwal S, Ghanty P, Pal NR. Identification of a small set of plasma signalling proteins using neural network for prediction of Alzheimer's disease. *Bioinformatics* 2015;31:2505–13.

17. Lasch P, Drevinek M, Nattermann H, et al. Characterization of *Yersinia* using MALDI-TOF mass spectrometry and chemometrics. *Anal Chem* 2010;**82**:8464–75.
18. Tang W, Wan S, Yang Z, et al. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 2018;**34**:398–406.
19. Yang Q, Li B, Tang J, et al. Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform* 2019, doi: [10.1093/bib/bbz049](https://doi.org/10.1093/bib/bbz049).
20. Li YH, Yu CY, Li XX, 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.
21. Vandenkoornhuyse P, Quaiser A, Duhamel M, et al. The importance of the microbiome of the plant holobiont. *New Phytol* 2015;**206**:1196–206.
22. Fu J, Tang J, Wang Y, et al. Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front Pharmacol* 2018;**9**:681.
23. Zhu F, Li XX, Yang SY, et al. Clinical success of drug targets prospectively predicted by in silico study. *Trends Pharmacol Sci* 2018;**39**:229–31.
24. Huang Q, Yang L, Luo J, et al. SWATH enables precise label-free quantification on proteome scale. *Proteomics* 2015;**15**:1215–23.
25. Tang J, Zhang Y, Fu J, et al. Computational advances in the label-free quantification of cancer proteomics data. *Curr Pharm Des* 2018;**24**:3842–58.
26. Wang P, Fu T, Zhang X, et al. Differentiating physicochemical properties between NDRIs and sNRIs clinically important for the treatment of ADHD. *Biochim Biophys Acta Gen Subj* 2017;**1861**:2766–77.
27. Yang Q, Wang Y, Zhang S, et al. Biomarker discovery for immunotherapy of pituitary adenomas: enhanced robustness and prediction ability by modern computational tools. *Int J Mol Sci* 2019;**20**:151.
28. Zhang X, Ning Z, Mayne J, et al. In vitro metabolic labeling of intestinal microbiota for quantitative metaproteomics. *Anal Chem* 2016;**88**:6120–5.
29. Yu CY, Li XX, Yang H, et al. Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate. *Int J Mol Sci* 2018;**19**:183.
30. Wang P, Zhang X, Fu T, et al. Differentiating physicochemical properties between addictive and nonaddictive ADHD drugs revealed by molecular dynamics simulation studies. *ACS Chem Neurosci* 2017;**8**:1416–28.
31. Cheng K, Ning Z, Zhang X, et al. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* 2017;**5**:157.
32. Zhang Y, Ying JB, Hong JJ, et al. How does chirality determine the selective inhibition of histone deacetylase 6? A lesson from Trichostatin A enantiomers based on molecular dynamics. *ACS Chem Neurosci* 2019; doi: [10.1021/acscchem-neuro.8b00729](https://doi.org/10.1021/acscchem-neuro.8b00729).
33. Zheng G, Yang F, Fu T, et al. Computational characterization of the selective inhibition of human norepinephrine and serotonin transporters by an escitalopram scaffold. *Phys Chem Chem Phys* 2018;**20**:29513–27.
34. Statnikov A, Henaff M, Narendra V, et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 2013;**1**:11.
35. Mak TD, Laiakis EC, Goudarzi M, et al. Selective paired ion contrast analysis: a novel algorithm for analyzing postprocessed LC-MS metabolomics data possessing high experimental noise. *Anal Chem* 2015;**87**:3177–86.
36. Tsalik EL, Henao R, Nichols M, et al. Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci Transl Med* 2016;**8**:322ra11.
37. Kumar M, Kumar Rath S. Classification of microarray data using kernel fuzzy inference system. *Int Sch Res Notices* 2014;**2014**:769159.
38. Neumann U, Riemenschneider M, Sowa JP, et al. Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Min* 2016;**9**:36.
39. Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Syst Appl* 2014;**41**:2052–64.
40. Goh WW, Wong L. Evaluating feature-selection stability in next-generation proteomics. *J Bioinform Comput Biol* 2016;**14**:1650029.
41. Ramus C, Hovasse A, Marcellin M, et al. Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. *Data Brief* 2016;**6**:286–94.
42. Li M, Gray W, Zhang H, et al. Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J Proteome Res* 2010;**9**:4295–305.
43. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* 2006;**103**:5923–8.
44. Zou Q, Zeng JC, Cao LJ, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 2016;**173**:346–54.
45. Han Z, Xue W, Tao L, et al. Identification of key long non-coding RNAs in the pathology of Alzheimer's disease and their functions based on genome-wide associations study, microarray, and RNA-seq data. *J Alzheimers Dis* 2019;**68**:339–55.
46. Xia J, Sinelnikov IV, Han B, et al. MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Res* 2015;**43**:W251–7.
47. 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.
48. Zou Q, Wan S, Ju Y, et al. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst Biol* 2016;**10**:114.
49. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**:2507–17.
50. Alvarez D, Hornero R, Marcos JV, et al. Assessment of feature selection and classification approaches to enhance information from overnight oximetry in the context of apnea diagnosis. *Int J Neural Syst* 2013;**23**:1350020.
51. Alshawaqfeh M, Bashairah A, Serpedin E, et al. Consistent metagenomic biomarker detection via robust PCA. *Biol Direct* 2017;**12**:4.
52. Vizcaino JA, Csordas A, del-Toro N, et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* 2016;**44**:D447–56.
53. Tabb DL, Vega-Montoto L, Rudnick PA, et al. Repeatability and reproducibility in proteomic identifications by liquid

- chromatography–tandem mass spectrometry. *J Proteome Res* 2010;**9**:761–76.
54. Billoir E, Navratil V, Blaise BJ. Sample size calculation in metabolic phenotyping studies. *Brief Bioinform* 2015;**16**:813–9.
 55. Tilocca B, Burbach K, Heyer CME, et al. Dietary changes in nutritional studies shape the structural and functional composition of the pigs' fecal microbiome—from days to weeks. *Microbiome* 2017;**5**:144.
 56. Mottawea W, Chiang CK, Muhlbauer M, et al. Altered intestinal microbiota–host mitochondria crosstalk in new onset Crohn's disease. *Nat Commun* 2016;**7**:13419.
 57. Schroeder BO, Birchenough GMH, Stahlman M, et al. Bifidobacteria or fiber protects against diet-induced microbiota-mediated colonic mucus deterioration. *Cell Host Microbe* 2018;**23**:27–40 e7.
 58. Pursiheimo A, Vehmas AP, Afzal S, et al. Optimization of statistical methods impact on quantitative proteomics data. *J Proteome Res* 2015;**14**:4118–26.
 59. Deutsch EW, Csordas A, Sun Z, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 2017;**45**:D1100–6.
 60. Xia J, Psychogios N, Young N, et al. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res* 2009;**37**:W652–60.
 61. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 2016;**11**:2301–19.
 62. Mathe EA, Patterson AD, Haznadar M, et al. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res* 2014;**74**:3259–70.
 63. Di Guida R, Engel J, Allwood JW, et al. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* 2016;**12**:93.
 64. Warrack BM, Hnatyshyn S, Ott KH, et al. Normalization strategies for metabolomic analysis of urine samples. *J Chromatogr B Analyt Technol Biomed Life Sci* 2009;**877**:547–52.
 65. Valikangas T, Suomi T, Elo LL. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief Bioinform* 2018;**19**:1344–55.
 66. Xue W, Yang F, Wang P, et al. What contributes to serotonin–norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem Neurosci* 2018;**9**:1128–40.
 67. Chai LE, Law CK, Mohamad MS, et al. Investigating the effects of imputation methods for modelling gene networks using a dynamic bayesian network from gene expression data. *Malays J Med Sci* 2014;**21**:20–7.
 68. Valikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* 2018;**19**:1–11.
 69. 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.
 70. Xue W, Wang P, Tu G, et al. Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder. *Phys Chem Chem Phys* 2018;**20**:6606–16.
 71. Fu T, Zheng G, Tu G, et al. Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations. *ACS Chem Neurosci* 2018;**9**:1492–502.
 72. McHugh ML. The chi-square test of independence. *Biochem Med* 2013;**23**:143–9.
 73. Varghese RS, Cheema A, Cheema P, et al. Analysis of LC-MS data for characterizing the metabolic changes in response to radiation. *J Proteome Res* 2010;**9**:2786–93.
 74. Farina D, Kamavuako EN, Wu J, et al. Entropy-based optimization of wavelet spatial filters. *IEEE Trans Biomed Eng* 2008;**55**:914–22.
 75. Hanna MH, Segar JL, Teesch LM, et al. Urinary metabolomic markers of aminoglycoside nephrotoxicity in newborn rats. *Pediatr Res* 2013;**73**:585–91.
 76. Fukushima A, Kusano M, Mejia RF, et al. Metabolomic characterization of knockout mutants in Arabidopsis: development of a metabolite profiling database for knockout mutants in Arabidopsis. *Plant Physiol* 2014;**165**:948–61.
 77. Westerhuis JA, van Velzen EJ, Hoefsloot HC, et al. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics* 2010;**6**:119–28.
 78. Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J* 2013;**4**:e201301009.
 79. Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle. *Brief Bioinform* 2013;**14**:315–26.
 80. Zhou L, Wang Q, Yin P, 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.
 81. Constantinou C, Chrysanthopoulos PK, Margarity M, et al. 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.
 82. Dutta B, Kanani H, Quackenbush J, et al. Time-series integrated 'omic' analyses to elucidate short-term stress-induced responses in plant liquid cultures. *Biotechnol Bioeng* 2009;**102**:264–79.
 83. Le Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* 2011;**12**:253.
 84. Ding Y, Wilkins D. Improving the performance of SVM-RFE to select genes in microarray data. *BMC Bioinformatics* 2006;**7**(Suppl 2):S12.
 85. Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon Rank-Sum test in small samples applied research. *J Clin Epidemiol* 1999;**52**:229–35.
 86. Kohl SM, Klein MS, Hochrein J, et al. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* 2012;**8**:146–60.
 87. Mischak H, Allmaier G, Apweiler R, et al. Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med* 2010;**2**:46ps2.
 88. Zhao Y, Hao Z, Zhao C, et al. A novel strategy for large-scale metabolomics study by calibrating gross and systematic errors in gas chromatography–mass spectrometry. *Anal Chem* 2016;**88**:2234–42.
 89. Li B, Tang J, Yang Q, 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.
90. Kim J, Mouw KW, Polak P, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* 2016;**48**:600–6.
 91. Tippmann S. Programming tools: adventures with R. *Nature* 2015;**517**:109–10.
 92. De Livera AM, Sysi-Aho M, Jacob L, et al. Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem* 2015;**87**:3606–15.
 93. Li B, Tang J, Yang Q, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* 2017;**45**:W162–70.
 94. Navarro P, Kuharev J, Gillet LC, et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* 2016;**34**:1130–6.
 95. Tyanova S, Albrechtsen R, Kronqvist P, et al. Proteomic maps of breast cancer subtypes. *Nat Commun* 2016;**7**:10259.
 96. Krawczuk J, Lukaszuk T. The feature selection bias problem in relation to high-dimensional gene data. *Artif Intell Med* 2016;**66**:63–71.
 97. Sarkar C, Cooley S, Srivastava J. Robust feature selection technique using rank aggregation. *Appl Artif Intell* 2014;**28**:243–57.
 98. Xia J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc* 2011;**6**:743–60.
 99. Feng W, Henning P, Wang MD. EGOMiner: a comprehensive genomics and proteomics data analysis and biological function interpretation system. *Conf Proc IEEE Eng Med Biol Soc* 2004;**4**:2809–12.