OPEN ACCESS
Freely Available Online

JOURNAL OF SYSTEMS BIOLOGY
ISSN NO: COMING SOON

Research Article

DOI : COMING SOON

Ovarian Cancer Identification Based on Feature Weighting for High-Throughput Mass Spectrometry Data

Lili Cui[1], Li Ge[1], Hongbin Gan[2], Xiaoping Liu[1, *], Yusen Zhang[1, *]

[1]School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China

[2]School of Marine Sciences, Shandong University at Weihai, Weihai 264209, China

Abstract

An important use of proteomics data from Mass Spectrometry (MS) is the classification of tumor types with respect to peptides in specific cancer types. It is highly critical to find an optimal set of markers among specific cancer peptides whose expression can be clinically utilized to build assays for the diagnosis or to track the progression of specific cancer types. A number of feature selection algorithms have been proposed to obtain the classification of MS data.

In this article, we proposed an improved feature selection algorithm based on feature weighting. Relief algorithm can calculate the weight of different features according to the correlation between their characteristics and categories. F-score is a simple filter-based feature selection method by evaluating how two sets of real numbers discriminate from each other. The main goal of this paper is to introduce a new feature weighting selection algorithm combining score from f-value and weight from relief, which is more accurate when classifying high-resolution MALDI-TOF (matrix-assisted laser desorption and ionization time-of-flight) MS data.

We have developed a four-step strategy for data processing based on: (1) Align the study sets by binning of raw MS data, (2) local maximum search(LMS) peak detection, (3) a new combination feature weighting selection algorithm and (4) support vector machines achieve a satisfactory performance of identifying cancer and the healthy. The best parameter set for LMS were achieved with control variable method, which achieve an average accuracy of 97.4167% (sd = 0.0146) and the best accuracy of 98.6111% in 1000 independent 10 -fold cross validations.

Corresponding Author: Yusen Zhang (zhangys@sdu.edu.cn), Xiaoping Liu (xpliu@sdu.edu.cn), School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China.

Keywords: Peak detection algorithm; Local Maximum Search; Feature weighting selection; Binning; F-score; Relief

Received: Jan 17, 2018 **Accepted:** Feb 25, 2018 **Published:** Mar 02, 2018

Editor: Kuei-Hu Chang, Department of Management Sciences, R.O.C. Military Academy, Taiwan.

Introduction

In the past 30 years, we have made great progress in the understanding and treatment of ovarian cancer. However, the overall 5-year survival rate of ovarian cancer patients is still hovering around 30%. Even though a research [1] shows that elderly patients with ovarian cancer mortality rates are generally higher than in younger patients. Ovarian cancer survival rates vary dramatically by stage. Within the stage, however, there are differences in survival by age, with younger women surviving better than older women even after adjustment for the general life expectancy of each age group (relative survival). For Stages 111-IV disease, women under 45 years of age have a 5-year relative survival rate of over 45% compared to only 8% for those 85 years of age and over [1]. In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype, i.e. observable trait. The genetic code stored in DNA is interpreted by mRNA, and the properties of gene expression give rise to the organism's phenotype. Such phenotypes are often expressed by the synthesis of proteins that control the organism's shape, or that act as enzymes catalyzing specific metabolic pathways characterizing the organism.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as transfer RNA or small nuclear RNA genes, the product is a functional RNA. The process of gene expression is used by all known life - eukaryotes, prokaryotes, and utilized by viruses - to generate the macromolecular machinery for life.

In eukaryotes, most mature RNA must be exported to the cytoplasm from the nucleus. While some RNAs function in the nucleus, many RNAs [2] are transported through the nuclear pores and into the cytosol, from where ribosome translating messenger RNA to chain of amino acids. For non-coding RNA the mature RNA is the final gene product. [3] In the case of messenger RNA the RNA is an information carrier coding for the synthesis of one or more proteins. Each mRNA molecule is translated into many protein molecules, on average ~2800 in mammals [4, 5].

There are two basic types of genetic mutations [6]: Acquired mutations are the most common cause of cancer. These occur from damage to genes during a person’s life. They are not passed from parent to child. Germline mutations, which are less common, are passed directly from a parent to a child. Mutations happen often, and the human body is normally able to correct most of them. Depending on where in the gene the change occurs, a mutation may be beneficial, harmful, or make no difference at all. So, one mutation alone is unlikely to lead to cancer. Usually, it takes multiple mutations over a lifetime to cause cancer. This is why cancer occurs more often in older people who have had more opportunities for mutations to build up [6].

Oncogenes turn a healthy cell into a cancerous cell [7]. Mutations in these genes are not inherited. Despite all that is known about the different ways cancer genes work, many cancers cannot be linked to a specific gene. Cancer likely involves multiple gene mutations [8]. Some evidence also suggests that genes interact with their environment [9], further complicating genes’ role in cancer. Doctors hope to continue learning more about how genetic changes affect the development of cancer. This knowledge may lead to improvements in finding and treating cancer, as well as predicting a person’s risk of cancer.

Ovarian cancer is a leading cause of many deaths, yet the biological sides of the disease and conventional measures of its detection are absent. We have used protein microarrays and autoantibodies from cancer patients to identify proteins that are aberrantly expressed in the ovarian tissue. Identifying proteins that reveal differences in the stages of neoplastic differentiation will be informative in understanding the disease. They may be useful for diagnostics and may also suggest useful targets for therapeutic intervention.

The novel biotechnology of high-throughput and high-resolution MALDI-TOF (matrix-assisted laser desorption and ionization time-of-flight) MS makes it promising to explore the low-molecular-weight (LMW) region of the blood proteome for the diagnosis of significant patterns for various diseases. In this work we considered the SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) high-resolution raw MS data provided by National Cancer Institute (NCI), on a study conducted to discriminate ovarian cancer from normal tissue. The published high-resolution data achieved with extensive quality control and assurance (QC/QA) analysis allow superior classification patterns when compared to those obtained with low-resolution instrumentation.

Methods

Binning

In the first step, we bin perform binning on raw MS data. Since the length of the observed m/z sequence varies in the raw MS data, align the study sets according to the sorted union of m/z ratios into an intensity frame with missing data. Usually, ensure that the individual spectra are well calibrated and, if necessary, use interpolation to put all spectra on the same timescale. Throughout this paper, the missing data are ignored in binning.

These results in a number of 'peak bins' across spectra. Specify an appropriate bin width to reduce the chance of nearby peaks being incorrectly coalesced into the same peak group. We label each unique peak group by the m/z value at the midpoint of its peak bin. Quantify the peaks for each individual spectrum using the maximum log intensity within each peak group. This method finds the significant local maxima in the spectrum and identifies an interval containing each peak. We label the peaks by the m/z value of the local maximum in the spectrum. Quantify each peak using the maximum log intensity on the individual spectra within the interval defining the peak on the spectrum. This approach allows the peak quantification to be robust enough to slight misalignment across spectra. This may seem surprising at first, but there is a good reason why this works. A peak is something that stands out above the noise and above the baseline should be preserved in the binned spectrum. The presence of the baseline does not affect our ability to detect the peaks. The success of the proposed method depends to an extent on having the spectra reasonably well binned at the beginning.

We binned the frame, at a given bin length l, into a matrix A of m-by-n, where n = 216(121 ovarian cancer samples and 95 control samples) and m is determined by l. Each bin is an interval of the form [b, b+l]. For the binned data, without ambiguity, the m/z ratio stands for the left boundary of an interval. After inquiry, binning with l = 0.42, has the most favorable performance in the next selection and classification, so the dimension is reduced from 373 401 to 26643.

Peak Detection and Qualification

Therefore, it is important to select features that are used for the identification of diseases in improving the classification accuracy and reducing the dimensionality of the dataset [10]. In an effort to choose the optimal subset of the predictor, different methods are employed. Feature extraction process is also an important part of pattern recognition and machine learning [11], including calibration of the spectra, baseline correction, normalization and denoising. Thanks to feature extraction process, the computation cost decreases and the classification performance can increase. It has been shown that the use of inadequate or ineffective methods in feature extraction may make it difficult to extract meaningful biological information from these data [12]. Peak detection is not only a feature extraction step, but also an indispensable step for subsequent protein identification, quantification and discovery of disease-related biomarkers [13, 14].

In this paper, we use LMS to denote Local Maximum Search. LMS is designed for single spectrum peak detection. This peak detection algorithm is designed according to Yasui's standards [14]. Local maximum means a peak, which is a local maximum of N neighboring points. It is desirable to remove baseline [15] and smoothing before peak detection [16]. During the peak detection, we use Yasui's standard: peaks should be the highest in the local neighborhood, which was defined by the parameter of the local neighborhood of the raw spectrum, and peak should be higher than the background at this point in the smoothed version.

Figure 1 gives a concrete example of peak detection by showing the result after each step of the peak detection process. We noted that smoothing and baseline correction may switch their locations in the pipeline. As shown in Figure 1, we get the final peak detection results through these steps. And the next step is to find the m/z ratios corresponding to the peak, and get the peak location and the peak value of the raw data for later use.

Feature Weighting Selection Algorithm

Feature selection refers to select a subset of M features from a set of N features, where M<N, such that the value of a criterion function is optimized over all subsets of size M. Ideally, feature selection methods search through the subsets of features, trying to find the best subset without losing the accuracy of the classification.

In this article, we proposed an improved feature selection algorithm based on feature weighting. We introduce a new feature weighting selection algorithm combine score from f-value with weight from relief in this paper, we call it f-value and relief feature weighting selection algorithm(FRFW).

Relief algorithm is a feature weighting algorithm, it can calculate the weight of each feature of the sample, give the weight of different features according to the correlation between the characteristics and categories.

F-score is a simple and effective algorithm including variable ranking as a principal selection mechanism. The larger the F-score is, the more likely the feature is more significant. In other words, a high F value (leading to a significant p-value depending on your alpha) means that at least one of your groups is much different from the rest, but it doesn't tell you

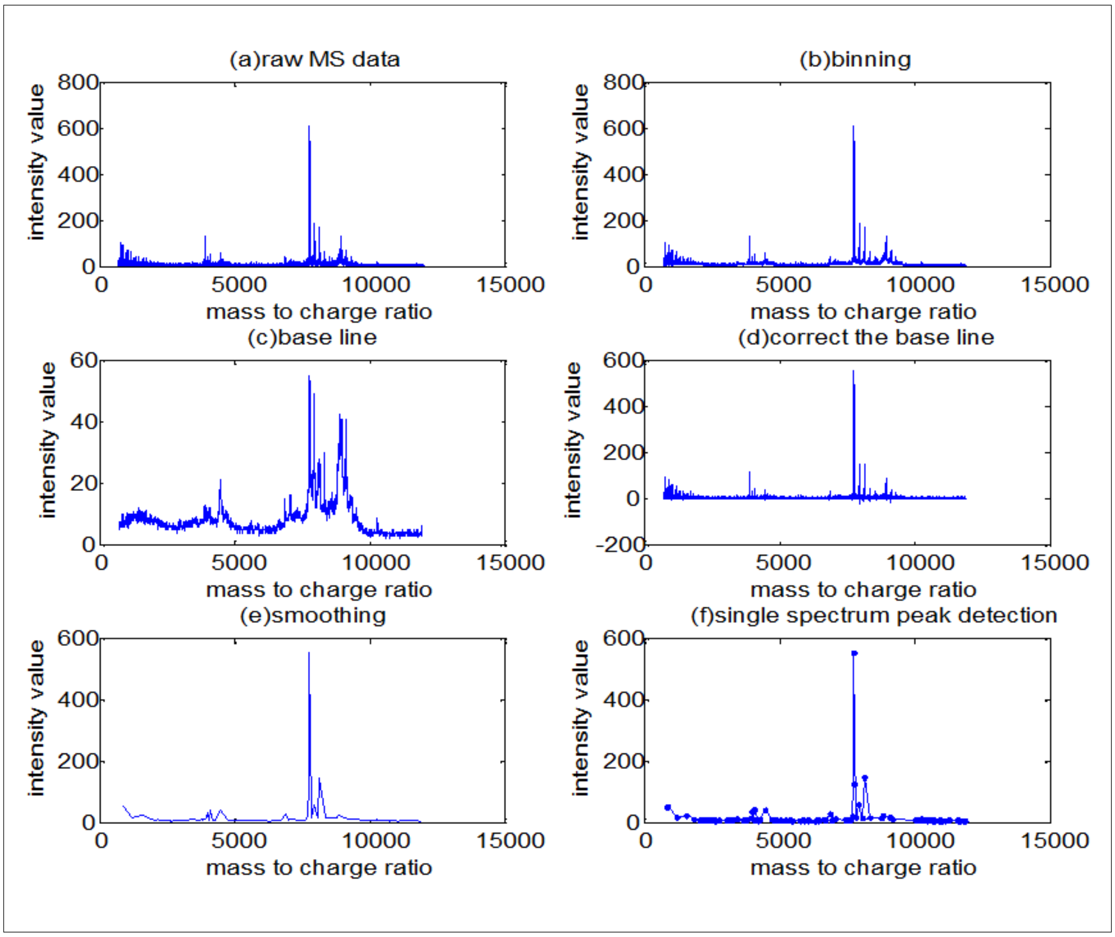

Figure 1. (a) A raw spectrum (b) align the study sets by binning of raw MS data (c) The minimum was detected as baseline in binning data (d) the spectrum after baseline correction (e) the spectrum after smoothing and baseline correction and (f) final peak detection and qualification results with peaks marked as points.

which group. Typically, one selects features that return high F-values and use those for further analysis.

F-score is used in feature selection to measure the discrimination of two sets of real-numbers. It reveals the discriminative power of each feature independently from others.

Traditional Dimensionality Reduction

Relief algorithm

Kira et al. [17] proposed the Relief algorithm in 1992. The Relief algorithm ranks the features according to it's the highest correlation with the observed class while taking into account the distances between opposite classes [18]. The main idea of the Relief algorithm is to estimate the quality of the features. The quality of the features is determined by their abilities to distinguish between observations those are closely related to each other.

The method borrows the idea of nearest neighbor learning algorithm. Relief selects m samples randomly from the training set. For the selected samples, two of their nearest neighbor samples are constructed, one contains the samples from similar class, another contains the samples from different class. To compare the selected samples and the two nearest neighbor samples, the correlation between each feature and class of each sample is obtained. Then the average value is used as the weight of each feature, and the correlation between each feature and class is obtained. Relief is only applicable to the case of two types of training samples. The algorithm averages the contribution of all hits and misses [19].

F-score

F-score is a simple feature selection filter method by evaluating the discrimination of two sets of real numbers. F-score is a simple and effective algorithm including variable ranking as a principal selection mechanism. The larger the F-score is, the more likely the feature is more significant. However, F-score cannot effectively reveal mutual information among features. For example, even if both features have low F-score, they can also be classified into two categories. Despite this disadvantage, F-score is effective and generally used with classifiers such as Support Vector Machine for accelerating the training and classification stage [20].

SVM Recursive Feature Elimination

SVMRFE is a well-known wrapper method for feature selection proposed by Guyon et al. [21], which defines the best feature set by using the Support Vector Machine (SVM).

SVM network is able to solve selection problem in the form of recursive feature elimination (SVM-RFE). In this approach, we use the SVM network with linear kernel. The idea of SVMRFE is that the orientation of the separating hyper-plane found by the SVM can be used to select informative features. If the plane is orthogonal to a particular feature dimension, then that feature is informative, and vice versa. In SVM-RFE approach to feature selection, we can eliminate irrelevant features step by step according to the assumed criterion related to their support in the discrimination of the classes. The SVM is retrained using smaller and smaller population of features. In each step, we eliminatethe features associated with the smallest absolute weights [22].

Kolmogorov–Smirnov Test

Kolmogorov–Smirnov test compares the medians of the groups of data to determine if the samples come from the same population. The null hypothesis is that both classes are drawn from the same continuous distribution. The alternative hypothesis is that they are drawn from different distributions.

In this paper we use two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test, alpha is a value between 0 and 1 specifying the significance level in Kolmogorov–Smirnov test, usually default is 0.05 for 5% significance.

Restriction of Coefficient of Variation

For a positive random variable X, the coefficient of variation (CV) is defined as c = sd(X)/E(X), which can be estimated by c^ = s/X¯ where s and X¯ are the sample standard deviation and sample mean respectively. The m/z ratio with relatively small CV is considered as a useful feature for the classification. The CV of intensity for the healthy and cancerous will be considered separately. Given CV thresholds of intensity, for instance with different tH and tC for the healthy and cancerous, the consequential dimensions of feature space are reduced by the restriction of the given coefficient of variation [23].

Support Vector Machine Learning Classifier

Next We use Support Vector Machine (SVM) for classification. The SVM [24] method is a widely used classification method of Statistical Learning Theory, originally started by Vapnik and Chervonenkis in the 1960s.

In case that the training set is linearly separable, the support vector classifier is the hyperplane with the maximal margin separating the two classified subsamples of the training set.

Generally, in the linearly non-separable case, we reach at a soft margin allowing training errors, where the classifier H is the solution of the optimization problem that is solved by the method of quadratic programming. Another approach to the linearly non-separable case is the kernel method. It constructs an optimal hyperplane decision function in feature space that is mapped from the original input space by using kernels, the training data are mapped into a higher dimensional feature space and become more separable.

Three types of commonly used kernel functions are:

Linear Kernel $k(\mathrm{x}i;\ \mathrm{x}j) = \mathrm{x}i\bullet\mathrm{x}j$

Polynomical Kernel $k(\mathrm{x}i;\ \mathrm{x}j) = (1 + \mathrm{x}i\bullet\mathrm{x}j)p$

Gaussian Kernel $k(\mathrm{x}i;\ \mathrm{x}j) = \exp(-||\mathrm{x}i - \mathrm{x}j||2/2\sigma 2)$

Data Set and Experiments

The Ovarian Dataset used in this experiment is serum proteomic data of ovarian cancer, it was obtained by the United States Food and Drug Administration (FDA) and the National Cancer Institute (NCI) for the analysis of serum samples from ovarian cancer and normal human body. This data is generated using a non-randomized study set of ovarian cancers and control specimens on an ABI Qstar fitted with a SELDI-TOF (matrix-assisted laser desorption and ionization time-of-flight) mass spectrometry (MS) technology to collect data relative to critical unanswered questions in the field of proteomic profiling.

High resolution time-of-flight (TOF) mass spectrometry (MS) proteomics data set from surface-enhanced laser/desorption ionization (SELDI) Protein Chip arrays on 121 ovarian cancer cases and 95 controls. The data sources can be accessed by FDA-NCI Clinical Proteomics at http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

We write the MS dataset as S = {(xi , yi)|xi ∈ Rm, yi = ±1, i = 1, 2, . . . , n}, where xi is an intensity vector according to a sorted sequence of m/z ratios and yi is the class label of xi (−1 for the healthy, +1 for cancer). When the feature space is high-dimensional, feature selection becomes crucial as the first step towards pattern recognition. For the raw ovarian high-resolution SELDI-TOF dataset composed of 95 control samples and 121 cancer samples, the dimension of the original feature space is over 370 000. Mass spectrometry data matrix of control and cancer is shown in Table 1.

Assuming that one finds p peaks from n spectra, this yields a p × n matrix of 'protein expression levels'. In the experiment, two-thirds samples are randomly chosen for training, and the remaining one-third samples are tested. The training set consisted of 64 normal samples and 80 cancer samples, and the remaining 31

Table 1. Mass spectrometry data matrix of control and cancer

m/z	−1	···	−1	+1	···	+1
r1	*X1,1*	···	*X1,k*	*X1,k+1*	…	*X1,n*
r2	*X2,1*	…	*X2,k*	*X2,k+1*	…	*X2,n*
···						
rm	*Xm,1*	…	*Xm,k*	*Xm,k+1*	…	*Xm,n*
	X1	…	*Xk*	*Xk+1*	…	*Xn*

For the sake of following data mining, the intensity observation is recorded in the column vector.

normal cases and 41 cancer cases form the testing set. The LMS peak detection in the new binned MS dataset obtained the total 371 valuable peaks from 216 spectras with the best parameter set, so the training set is 144 x 371 dimensional matrix and the testing set is 72 x 371 dimensional matrix.

And then the FRFW algorithm, which combine score from f-value with weight from relief, introduced in this paper is used to feature selection and feature extraction. In this article, to combine score from f-value with weight from relief, we need to normalize the score from f-value and the weight from relief to represent the data on a systematic scale. After normalization, two normalized values are added as the joint weight of the feature. The new feature weighting selection algorithm is a simple and effective algorithm including variable ranking as a principal selection mechanism. It can calculate the joint weight of each feature of the sample, the larger the joint weight is, the more likely the feature is more significant.

However, the method selects subset features depend on the parameter sets of Local Maximum Search peak detection algorithm. The LMS peak detection algorithm usually requires parameter optimization to obtain accurate results when it performs on a different type of data sets, change each parameter of Local Maximum Search peak detection algorithm will affect the selection. The parameters of Local Maximum Search peak detection algorithm consist of baseline correction method, baseline correction window size, peak width constraint parameter and neighbor size. In the parameter inquiry, the control variable method is adopted. Table 2, together with Figure 2, we can see baseline correction method has a major impact on the classification accuracy, use the mean of the points as the baseline points lead to a promising application prospect.

The FRFW algorithm proposed in this paper is combined score from f-value with weight from relief, comparison results of FRFW algorithm with relief algorithm and the f-score algorithm is shown in sub-figure (d) of Figure 2. Table 3 gives comparison results of FRFW algorithm with other feature selection algorithms, FRFW algorithm has the best classification accuracy. In summary, we get an intuitive and accurate conclusion, feature selection has a strong dependence on peak extraction, and LMS method is an effective tool for cancer type prediction of peptide markers. As is shown in Table 2, the best parameter set for LMS were achieved while the window size for baseline correction is 11, the peak width constraint parameter is 9, and use the minimum of the points as the baseline point within 48 local neighborhoods.

It should be noted that in feature selection, there is a great relationship between data types and feature selection algorithms. Different feature selection algorithm was used to obtain accurate results when it performs on a different type of data sets. For example, relief algorithm makes a good showing in ovarian low-resolution MALDI-MS data and SVM recursive feature elimination do better in ovarian high-resolution SELDI-TOF data when used alone. Nevertheless, the performance of our FRFW algorithm was better than the other algorithms reported in the literature and classifiers found in data-mining of the ovarian high-resolution SELDI-TOF data set.

At the same time, in order to illustrate the superiority of the new classification models proposed, some other traditional dimensionality reduction algorithms are compared in the paper. After binning, the feature dimensions are up to 26643, to address the "curse of dimensionality" problem, these dimensionality reduction algorithms can be used before peak detection. Due to the particularity of the high-resolution MALDI-MS data, they were not well received. We got the consequential dimensions of feature space and testing accuracy after a CV restriction or KS-test, the results are shown in table 4 and table 5. They were derived from best training with the use of Local Maximum Search peak detection algorithm with the best parameter set, new feature weighting selection algorithm and support vector machine learning classifier. Even though Kolmogorov–Smirnov test has a very high consistency with other feature selection algorithms, such as Relief algorithm, SVM recursive feature elimination and f-score algorithm, in feature selection of ovarian low-resolution MALDI-MS data, it has a not so good application prospect for feature reduction of ovarian high-resolution SELDI-TOF data, and so does CV restriction.

To compare, we conducted a comprehensive

Table 2. Local Maximum Search peak detection algorithm parameter sets and testing accuracy (mean ± standard error, %) with classification models derived from best training, with the use of new feature weighting selection algorithm and support vector machine learning classifier, the best result of the data set is highlighted in bold.

Local Maximum Search peak detection algorithm parameter				Expected testing accuracy	Best testing accuracy
baseline correction method	baseline correction window size	peak width constraint parameter	neighbor size		
Mean	**11**	**9**	**48**	**97.4167 ± 1.4583**	**98.6111**
Median	5	6	66	96.8056 ± 0.7015	98.6111
min	8	6	57	95.3056 ± 0.9676	97.2222

Table 3. Expected testing accuracy (mean ± standard error, %) and best testing accuracy with classification models derived from best training, with the use of different feature selection algorithm, the best result of the data set is highlighted in bold.

feature selection algorithm	Expected testing accuracy	Best testing accuracy
FRFW	**97.4167 ± 1.4583**	**98.6111**
Relief	95.2222 ± 0.6964	95.8333
F-score	96.0278 ± 0.6281	97.2222
SVM-rfe	96.4722 ± 0.6992	97.2222
KS test	95.6111 ± 0.5143	95.8333
Restriction of CV	93.3889 ± 1.3029	95.8333

Table 4. The consequential dimensions of feature space and testing accuracy (mean ± standard error, %) after a CV restriction, with classification models derived from best training, the best result of the data set is highlighted in bold.

CV restriction	Feature dimensions	Expected testing accuracy	Best testing accuracy
tH = 0.3 & tC = 0.3	16102	82.3611 ± 0.9412	83.3333
tH = 0.35 & tC = 0.35	23435	90.1667 ± 1.1852	91.6667
tH = 0.4 & tC = 0.4	**25148**	**93.3889 ± 1.3029**	**95.8333**
tH = 0.45 & tC = 0.45	26040	92.5556 ± 1.8773	95.8333
tH = 0.5 & tC = 0.5	26381	92.3889 ± 0.8529	94.4444

Table 5. The consequential dimensions of feature space and testing accuracy (mean ± standard error, %) after KS-test with different p-values, with classification models derived

KS-test with different p-values	Feature dimensions	Expected testing accuracy	Best testing accuracy
0.0500	13954	92.1111 ± 0.8152	93.0556
0.0100	10509	94.5556 ± 0.4729	95.8333
0.0050	**9249**	**95.6111 ± 0.5143**	**95.8333**
0.0010	6834	92.4722 ± 1.4332	94.4444
0.0005	5969	93.6667 ± 0.6964	94.4444
0.0001	4449	91.6667 ± 1.0117	93.0556

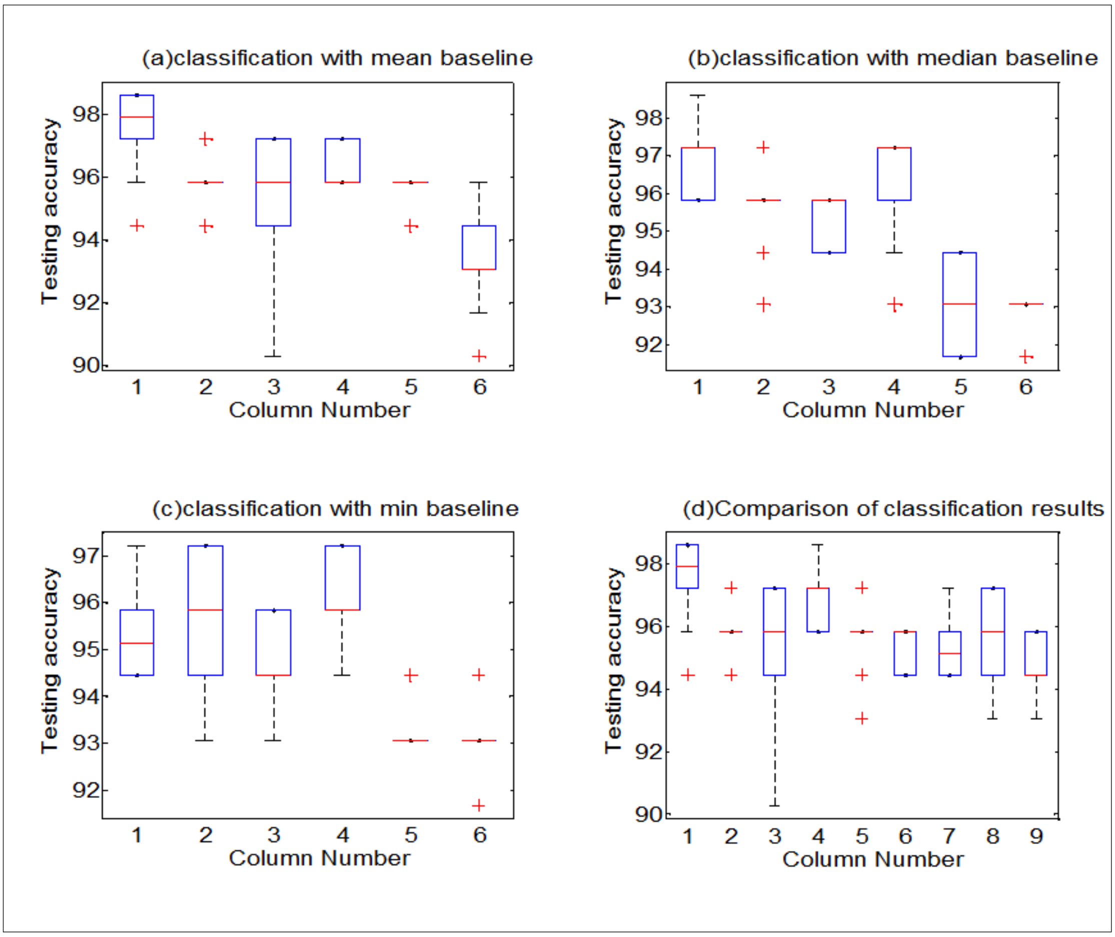

Figure 2. Average testing accuracy with classification models derived from best training. In sub-figure(a)(b)(c), the results shown in column 1 to column 6 are obtained by using FRFW, Relief, F-score, SVM-rfe, KS test, Restriction of CV, respectively. In sub-figure(d), the results shown in column 1 to column 9 are obtained by using FRFW, Relief, F-score with mean (column 1 to 3), median (column 4 to 6), min (column 7 to 9) baseline, respectively.

experimental study using high-resolution MALDI-MS data. In the case of peak detection algorithms, Local Maximum Search peak detection algorithm provides a good performance. Regarding feature selection, as shown in Table 3, when classification models derived from the best training were compared, the FRFW algorithm outperformed in classification. As for the learning classifiers, SVMs performed the best with respect to the expected testing accuracy. Consequently, we developed a four-step strategy for data processing based on: (1) Align the study sets by binning of raw MS data, (2) local maximum search(LMS) peak detection, (3) the FRFW algorithm and (4) support vector machines achieve a satisfying performance of identifying cancer and the healthy.

Serum proteomic profiling is a new approach to cancer diagnosis. However, it confronts a challenging environment, as it combines measurement technologies that are new in the clinical setting with novel approaches to processing and interpreting high dimensional data. Further, controlling large clinical studies can be challenging even in more established settings. Nevertheless, it represents an advance in the ability to diagnose and understand the illness.

The classifier-independent data preprocessing of proteomic MS data shows a promising approach to the coming classification. Since the issue of different feature selection methods and different classification models as they relate to classification performance has not been addressed, more robust classifiers are still urgently needed, as well as their ensemble. In addition, the precisions could be further improved by some resampling method [25], which assigns every testing sample point a probability of being cancer.

References

1. Ries, Lynn A. Gloeckler. "Ovarian cancer: survival and treatment differences by age." Cancer 71.S2 (1993): 524-529.
2. Köhler A, Hurt E (October 2007). "Exporting RNA from the nucleus to the cytoplasm". Nat. Rev. Mol. Cell Biol. 8(10): 761–73.
3. Amaral PP, Dinger ME, Mercer TR, Mattick JS (March 2008). "The eukaryotic genome as an RNA machine". Science 319 (5871): 1787–9.
4. Schwanhäusser B, Busse D, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011). "Global quantification of mammalian gene expression control". Nature 473 (7347): 337–42.
5. Schwanhäusser B, Busse D, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2013). "Corrigendum: Global quantification of mammalian gene expression control". Nature 495 (7439): 126–7.
6. Pleasance, Erin D., et al. "A comprehensive catalogue of somatic mutations from a human cancer genome." Nature 463.7278 (2010): 191-196.
7. Weinberg, Robert A. "Oncogenes, antioncogenes, and the molecular bases of multistep carcinogenesis." Cancer Research 49.14 (1989): 3713-3721.
8. Loeb, Lawrence A., Keith R. Loeb, and Jon P. Anderson. "Multiple mutations and cancer." Proceedings of the National Academy of Sciences 100.3 (2003): 776-781.
9. Armstrong, Bruce, and Richard Doll. "Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices." International journal of cancer 15.4 (1975): 617-631.
10. Wang Y, Tetko IV, Hall MA (2005) Gene selection from microarray data for cancer classification a machine learning approach. Comp Biol Chem 29: 37–46.
11. Ani, A. Al. (2005). Feature subset selection using ant colony optimization. International Journal of Computational Intelligence, 2(1), 53-58.
12. Morris J S, Coombes K R, Koomen J, et al. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum [J]. Bioinformatics, 2005,21 (9):1764-1775.[13] Gras R, Müller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF, Appel RD: Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 1999, 20:3535-3550.
13. Gras R, Müller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF, Appel RD: Improving protein

identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 1999, 20:3535-3550.

14. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL: Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Research 2002, 62:3609-3614.

15. Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, Tracy ER, Trosset MW, Sasinowski M, Semmes OJ, Manos DM: Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. Clinical Chemistry 2005, 51:65-74

16. Yang C, He Z, Yu W: Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics 2009, 10:4.

17. KIRA K, RENDELL L A. The features selection problem: traditional methods and new algorithm[C]/ Proc of the 10th National Conference on Artificial Intelligence. Michigan: AAAI Press,1992:129-134.

18. Robnik-Sikonja, R., & Kononenko, I. (2003). Theoretical and empirical analysis of Relief and RRelief. Machine Learning, 53, 23–69.

19. Tomasz Latkowski, Stanislaw Osowski: Data mining for feature selection in gene expression autism data. Expert Systems with Applications, 2015,42:864–872

20. Sheng Ding: Feature Selection based F-score and ACO Algorithm in Support Vector Machine. 2009 Second International Symposium on Knowledge Acquisition and Modeling, KAM.2009.137

21. Guyon I, Weston J, Barnhill S, Vapnik VN: Gene selection for cancer classification using support vector machines. Machine Learning 2002, 46(1–3):389-422.

22. Qingzhong Liu, Andrew H Sung, et al. Comparison of feature selection and classification for MALDI-MS data. BMC Genomics 2009, 10 (Suppl 1): S3

23. Yu J S, Ongarello S, Fiedler R, et al. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data[J]. Bioinformatics, 2005,21 (10): 2200-2209.

24. Vapnik, V.N. (1998) Statistical Learning Theory. John Wiley & Son, Inc., New York

25. Gelman, A., Carlin, J.B., Stern, H.S. and Rubin,D.B. (2004) Bayesian Data Analysis, 2nd end. Chapman & Hall/CRC Press.