Engineering, **2013, 5, 433-439**
http://dx.doi.org/10.4236/eng.2013.510B089 Published Online October 2013 (http://www.scirp.org/journal/eng)

A Localized-Statistic-Based Approach for Biomarker Identification of Omics Data

Kuan Zhang, He Chen, Yongtao Li
Beijing Aerospace Control Center, Beijing, China
Email: zhangkua@mail.ustc.edu.cn, chenhe_mail@netease.com, xlslyt@sina.com

Received 2013

ABSTRACT

Omics data provides an essential means for molecular biology and systems biology to capture the systematic properties of inner activities of cells. And one of the strongest challenge problems biological researchers have faced is to find the methods for discovering biomarkers for tracking the process of disease such as cancer. So some feature selection methods have been widely used to cope with discovering biomarkers problem. However omics data usually contains a large number of features, but a small number of samples and some omics data have a large range distribution, which make feature selection methods remains difficult to deal with omics data. In order to overcome the problems, we present a computing method called localized statistic of abundance distribution based on Gaussian window (LSADBGW) to test the significance of the feature. The experiments on three datasets including gene and protein datasets showed the accuracy and efficiency of LSADBGW for feature selection.

Keywords: Protein-Omics Data; Biomarker Selection; Localized Statistic; Gaussian Window

1. Introduction

With the advent of high-throughput measurement techniques such as transcriptome by microarray and proteome by mass spectrometry, the omics, which mean comprehensive analysis of a specific layer in a cellular system and are emerging as essential methodological approaches for molecular biology and systems biology, have been accumulated rapidly and make it possible to capture the entire snapshot of cell-wide activity [1,2]. The increase in data acquisition has lead to a demand for practical and effective data mining methods for in silico analysis. One of the strongest challenge problems biological researchers have faced is to find the methods for discovering biomarkers for tracking the process of disease such as cancer [3,4], as the biomarkers selection can be viewed as a major bottleneck of supervised learning and data mining on omics data [5,6].

Feature selection approaches, which aim to find a set of features that best discriminate biological samples of different types, have been widely applied to cope with discovering biomarkers problem [3,4,7-9]. The selected features are "biomarkers", and they form "marker panel" for analysis. The fold-change and p-value are two commonly known criteria to select differentially expressed features under two experimental conditions. In the fold-change method, a feature is viewed as a "biomarker" if the ratio in absolute value of the expression levels between two classes exceeds a certain threshold, e.g., a 2-fold change. The p-value ranking is an alternative approach for feature selection. Often the p-value is the probability outcome from a statistical testing procedure that there is no difference between two conditions for an individual feature. A variety of statistical tests including two-sample t test [10-16], X^2 test [10,17], the one-way analysis of variance [18,19], the Wilcoxon signed rank test [20-23] and Mann-Whitney test [23] have been used to obtain the p-values. Though great success have been obtained using these approaches in selecting biomarkers, it still remains difficult to deal with omics data. As we know that omics datasets always belong to small sample datasets, because the number of features significantly outnumbers the number of samples. Then the p-value methods based on statistical tests sometimes are failed to deal with the omics data, for example, if the sample number of the dataset only equals to 1 for each class the statistical tests miss their efficiency. And [24] indicates that some omics data have a large range distribution, so the same criteria for different range data which is the strategy employed by fold-change approach is incorrect, for example, the significance of 2-fold change from 2 to 1 is not equal to the significance of 2-fold change from 20,000 to 10,000.

In order to overcome the large range problem, [24] developed a computing method called Localized Statistics of Protein Abundance Distribution (LSPAD) to eva-

luate the statistical significance of protein-abundance bias between two classes, by which are differentia significance of a particular protein should be calculated through its local protein-abundance distribution-window rather than through whole distribution range from the lowest to highest protein abundances. In fact, even though the sample number of the dataset only equals to 1 for each class LSPAD also shows good performance which is validated in [24]. However LSPAD is under-utilized practice and there are two shortcomings in LSPAD. The first is that the strategy of selecting local distribution window is too rough, which postulated a width of the local window for statistics as 33%, *i.e*. only neighbored proteins within the 33% A-axis around a particular protein should be used for calculation. And the second is that LSPAD employs the fisher exact test to check the statistical significance. Fisher exact test is a statistical significance test used in the analysis of contingency tables where sample sizes are small. However if the data type is float rounding operation must be performed which may make Fisher exact test fail to deal with the omics data and fisher exact test should be time-consuming when the sample sizes are large.

In this study we present a computing method called localized statistic of abundance distribution based on Gaussian window(LSADBGW) which also employs the localized statistic strategy used by LSPAD but propose a Gaussian window as the local abundance distribution window and a simpler and more general statistic approach to test the significance of the feature. By using the Gaussian window, the selection of local abundance distribution window is more reasonable and persuasive. And LSADBGW not only can deal with the integral data but also the float data, which furthers the application range comparing with LSPAD. The experiments on three datasets including gene and protein datasets and the comparison with the LSPAD show the accuracy and efficiency of LSADBGW for feature selection.

In summary, our contributions are: 1) We extend the application range of localized statistic strategy to all omics, which is opposite to LSPAD is only oriented towards the protein tandem mass spectrometry data processed by SEQUEST [25]; 2) We propose a new strategy of selecting local abundance distribution window which employs the Gaussian window. By using the Gaussian window our method is more reasonable and persuasive than LSPAD; 3) We proposed a simpler but more effective statistic test instead of the fisher exact test used in LSPAD. The rest of the paper is organized as the follows. A brief not on the LSPAD is given in Section 2. Our method is presented in Section 3 and the datasets and experiments are given in Section 4. We show the experimental results and discuss the results in Section 5. Finally Section 6 concludes.

2. Related Work

The concept of localized statistic used in feature selection of omics is firstly proposed by [24], in which human serum of non-diabetic and diabetic cohorts was analyzed by proteomic approach. To analyze total 1377 high-confident serum-proteins, they developed a computing strategy called localized statistics of protein abundance distribution (LSPAD) to calculate a significant bias of a particular protein-abundance between these two cohorts.

The LSPAD method can be divided to two steps. Firstly, since the peptide-spectral-count distributions of identified serum-proteins were widely spread out to the range of 10^5, they developed *M-A* plotting referring to microarray analysis in order to display a relative protein-abundance distribution of each protein. The *M* and *A* values are defined as follows:

$$\begin{aligned} A &= (Y_1 + Y_2)/2 \\ M &= Y_1 - Y_2 \\ Y_1 &= \log_2(X_1 + 1) \\ Y_2 &= \log_2(X_2 + 1) \end{aligned} \tag{1}$$

wherein X_1 and X_2 respectively represent the peptide spectral counts in diabetic serum and in non-diabetic serum, *M* represents differential protein abundance between diabetic and non-diabetic serum, and A represents the average protein abundance.

Then the differential significance of a particular protein is calculated based on the proteins fell into its local protein-abundance distribution-window using fisher's exact test. And [24] postulates a width of the local window for statistics as 33% A-axis.

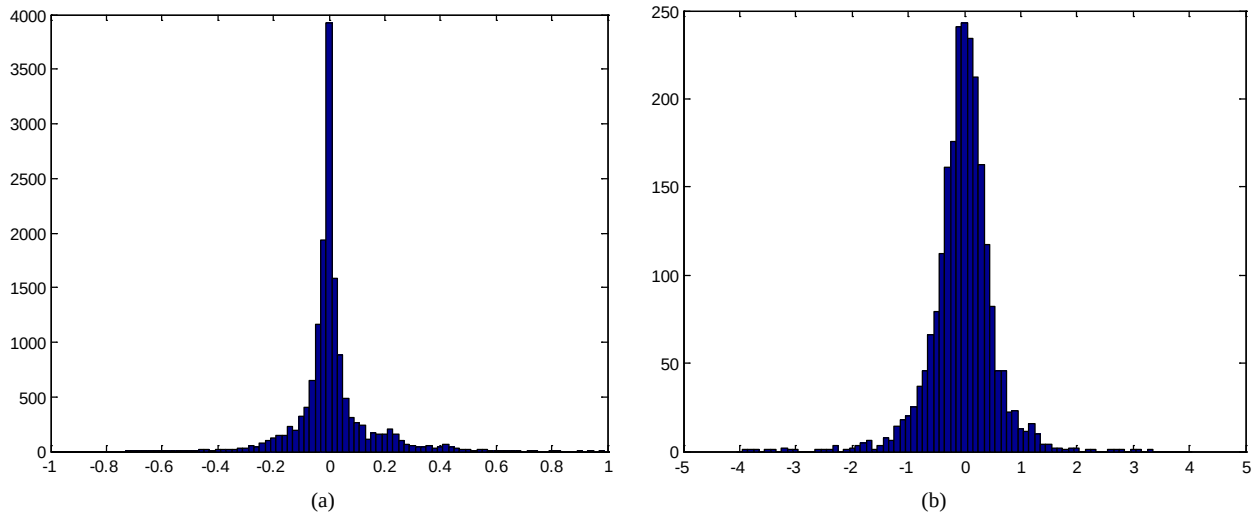
3. Method

In order to overcome the under-utilized in practice and the unreasonable window selection strategy, we proposed a more practical and reasonable method of selecting significant features called localized statistic of abundance distribution based on Gaussian window (LSADBGW). In fact, the M value used in Equation (1) can be employed as statistic value; on the contrary, M value is ignored by LSPAD. Because of the generality and simplicity of the normal distributions, it has been widely used in various areas, including the omics data such as gene expression data [26]. And we propose a Gaussian window in LSADBGW instead the local window used in LSPAD.

3.1. The Significant Test Method Using M Value

We assume that the *M* value obeys the normal distribution, and this is reasonable which can be validated in **Figure 1**.

With the assuming a Gaussian distribution, the signi-



(a) (b)

Figure 1. The M values distribution. (a) represents the M values distribution of serum SELDI MS data (Ovarian, 07 August 2002); (b) represents the M values distribution of wing sarcoma and rhabdomyosarcoma in the dataset small round blue cell tumors which is a DNA microarray dataset.

ficance of a feature can be given by

$$S = \frac{M^{data} - < M_W >}{\sigma_W} \tag{2}$$

wherein S represents the significance value, M^{data} represents the M value of the feature tested, "M_W" represents the mean value of the M values fell in the statistical window and σ_W represents the standard deviation of the M values fell in the statistical window.

After S value is obtained by Equation (2), the significance can be calculated through S e.g. $|S| \geqq 2.6$ can be treated as significant at a level of 99% assuming a Gaussian distribution.

3.2. The Gaussian Window

Since the Significance calculation of particular differential features should be localized to a certain range of related abundance level [24], the selection of appreciate local abundance distribution window plays an important role in localized statistics method. However choosing a local window for localized statistics appropriate to all kinds of data distribution, which ensures that all the data fell into it are under the same range, is difficult or impossible, as the concept of the same range is puzzled. Then we consider the interaction between different range samples instead of accurate the same range partition, that is, the correlation between samples located nearby with each other is higher than the samples located far. For example, under the data partition of [24], the correlation between low level and high level of protein abundance samples is lower between two high level samples.

However, how to accurately define and quantify the correlation between two samples according to their range distance is also a problem. Fortunately, it is known that there is close relationship between data range and data distribution, that is, the problem of estimating the correlation between two different range samples may be redefined and carried out from the view of the density estimation of distribution. So the correlation between two samples can be performed according to the contribution to the density estimation of each sample point for each other. For example, if sample point A has a higher contribution for the density estimation of sample C than the point B, we can say that the relationship between A and C is higher than A and B.

So from the point view of density estimation, the selection of location range window can employ the same strategy of location density estimation window. In fact, LSPAD employs rectangle window which the width is the 33% of all the range length. However this seems not reasonable, that is, it is difficult to say that using 33% is better than using 25% or others. We focus on the Gaussian window instead of rectangle window.

With a generalized weight kernel function $K(x)$ the density estimator $\hat{p}(x)$ is given by

$$\hat{p}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{3}$$

wherein N is the sample number, h is called smoothing parameter or window width and the kernel function $K(x)$ is required to be a normalized probability density. If $K(x)$ is the Gaussian kernel, the density estimator is given by

$$\hat{p}(x) = \frac{1}{N}\frac{1}{h\sqrt{2\pi}}\sum_{i=1}^{N} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right) \tag{4}$$

The choice of the bandwidth h is crucial to the density estimator, that is, if h is chosen to small spurious fine

structure becomes visible, while if *h* is too large all detail, spurious or otherwise is obscured. There are some methods for choosing an appropriate bandwidth available, however most of these methods suffer a considerable computational burden [27]. As a tradeoff between computational effort and performance one may choose the optimal bandwidth as the one that minimizes the mean integrated square error, assuming the underlying distribution is Gaussian. An optimal Gaussian bandwidth h_{opt} is given by [28]

$$h_{opt} = (\frac{4}{3N})^{\frac{1}{5}}\sigma \approx 1.06\sigma N^{-\frac{1}{5}} \quad (5)$$

We employ the Gaussian window as the local abundance distribution window. In fact the Gaussian window used is not the original local window, on the contrary, it is a whole window but the weight for each sample point is different. The sample set used to localized statistics is constructed by the follow strategy

$$\begin{aligned} &\textit{for each sampe } x_i \\ &\quad \textit{if } random_i < P(x_i) \\ &\quad\quad \textit{then select } x_i \textit{ to staDataset} \\ &\quad \textit{end if} \\ &\textit{endfor} \end{aligned} \quad (6)$$

wherein $randow_i$ is a random number obey the uniform distribution between 0 and 1, *staDataset* represents the sample set used to localized statistics and $P(x_i)$ is given by

$$p(x_i) = 2*(1 - normcdf(x, h_{opt}, |x_i|)) \quad (7)$$

wherein *norcdf* (*x*, *hopt*, $|x_i|$) is defined as the normal cumulative distribution function, *x* represents the mean of the normal distribution function, h_{opt} represents the standard deviation and $|x_i|$ means the absolute value of the sample x_i.

4. Materials and Experiments

4.1. Datasets

Three datasets are deployed here:

Dataset1: Ovarian cancer Dataset (07 August 2002), which was collected using WCX2 protein array. The sample set included 91 controls and 162 ovarian cancers. The SELDI MS data for each case is an ASCLL file containing 15,155 points of *m/z* values with corresponding intensities.

Dataset2: Small Round Blue Cell Tumors (SRBCTs), which was obtained from glass-slide cDNA microarrays. The data consisted of expression measurements on 6567 genes (2308 genes after filtering for minimal level of expression). The tumors are classified as Burkitt lymphoma (BL, 11 samples), Ewing sarcoma (EWS, 29 samples), neuroblastoma (NB, 18 samples) and rhabdomyosarcoma (RMS, 25 samples). As we only focus on the binary classification problem, EWS and RMS are selected to form a new two class dataset.

Dataset3: Stem Cell Matrix (SCM) [29], which is a database of global gene expression profiles. The database consisted of 218 samples which belong to 17 cell lines. As the operation in dataset2, ES cells_undifferentiated and ES_differentiated neural stem cells are selected to form a new two class dataset. IPS cells also are selected to further our method and this will be discussed in the latter section.

4.2. The Classification Results and Discussion

The LSADBGW currently is suitable for the two column data, so the mean vectors of two classes must be calculated firstly and form a new mean dataset. In fact, this operation may ignore the differences among the same class data which are useful for feature selection. Leave-one-out-cross-validation (LOOCV) method and liner-SVM are employed in our classification experimental framework.

As the mean vectors are only used for three methods, the differences between the same classes samples are ignored which may be an obstacle for classification. After the feature selection, we cluster the features selected to 10 classes by k-mean cluster method, and then we selected the top 1 feature of each class to form a feature sets for classification.

In **Figures 2**-**4** we respectively list the results obtained from the dataset 1, dataset 2 and dataset 3 using LSADBGW, LSPAD' and LSPAD. Here, all the p values used in three methods were equal to 0.95.

The results in **Figure 2** showed that the LSPAD performed better than LSPAD', which seems that the fisher' exact test was better than using simple statistical test using *M* values. However, in **Figured 3** and **4**, the results generated by LSPAD were not represented. This is because that the LSPAD did not generate good significant features set which were illustrated in **Figures 5** and **6**. The results in **Figure 5** showed that only two features were selected while in **Figure 6** showed that almost all the features were selected, this phenomena indicated that the LSPAD using the fisher' exact test was not a stable strategy for omics data, on the contrary, the LSPAD' using simple statistical test were much more stable.

It was also showed that the performance of using Gaussian window performed better than rectangle window, especially in **Figure 2**. However the results in **Figure 4**, LSPAD' seems a little better than LSADBGW. We then respectively used the top 10 and top 20 features without clustering to investigate the performance of LSADBGW and LSPAD', and the results were showed in **Figure 7**

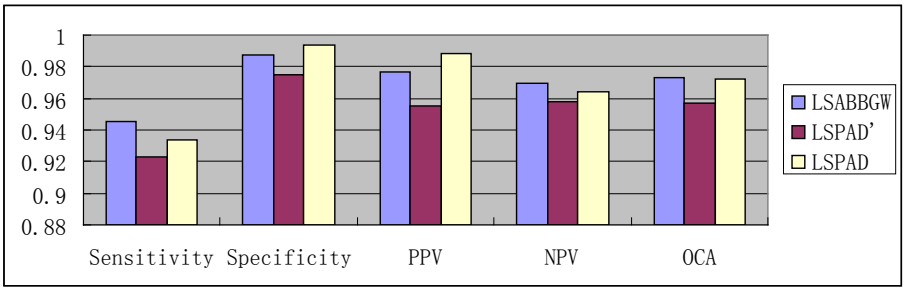


Figure 2. The classification performance comparison on the ovarian cancer dataset.

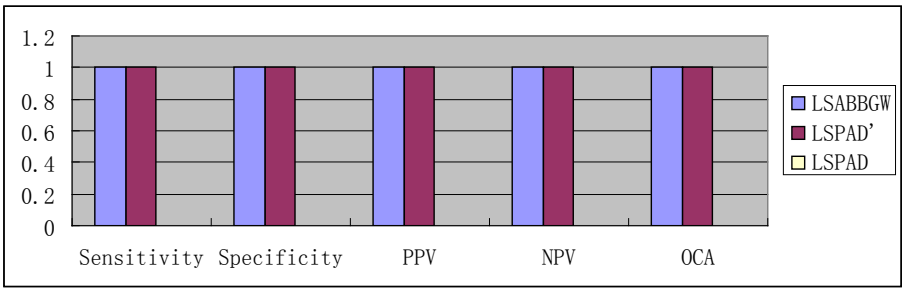


Figure 3. The classification performance comparison on the small round blue cell tumors dataset.

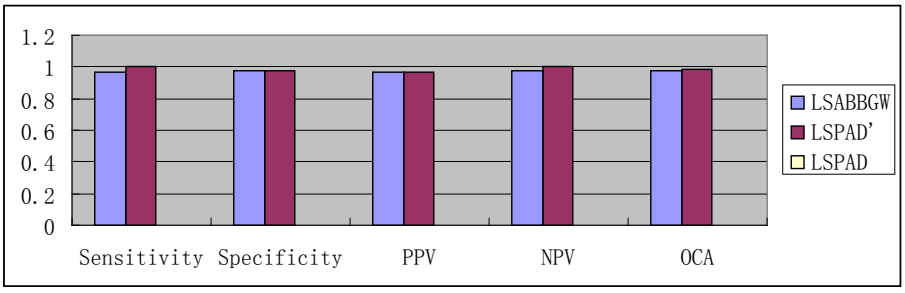


Figure 4. The classification performance comparison on the stem cell matrix dataset.

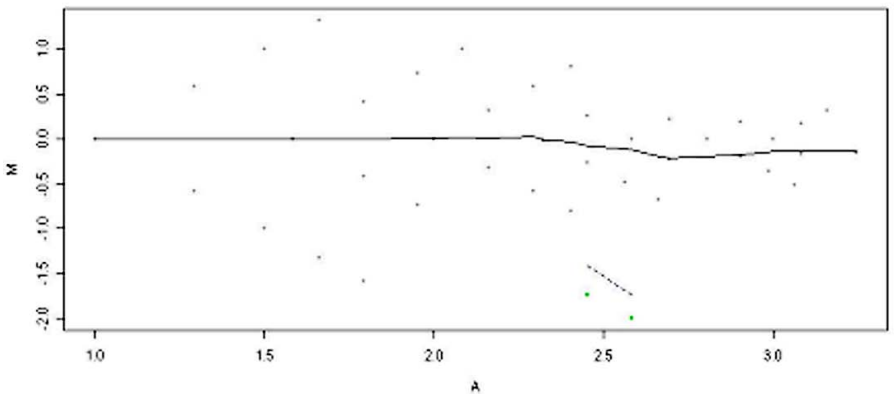


Figure 5. M-A plotting of small round blue cell tumors dataset, red dots represented statistically significant overpresented genes in EWS and Green dots represented statistically significant under-represented genes in EWS.

and 10. The new results, especially in **Figure 8**, indicated that the performance of LSADBGW was better than LSPAD', which meant that the strategy employing the Gaussian window performs better than employing the rectangle window.

The comparative study of three feature selection methods indicated that the strategy employing simple statistical test using *M* values was much more stable than fisher' exact test and employing the Gaussian window is much more accurate than rectangle window.

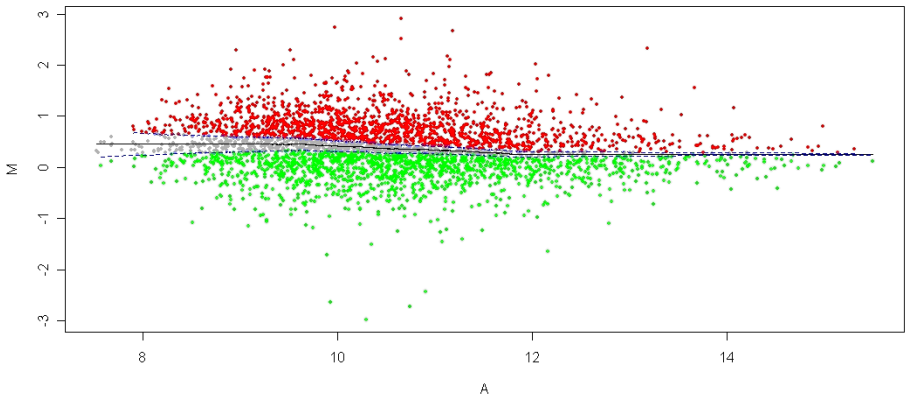


Figure 6. M-A plotting of stem cell matrix dataset, red dots represented statistically significant over-presented genes in ES cells_undifferentiated and green dots represented statistically significant under-represented genes in ES cells_undifferentiated.

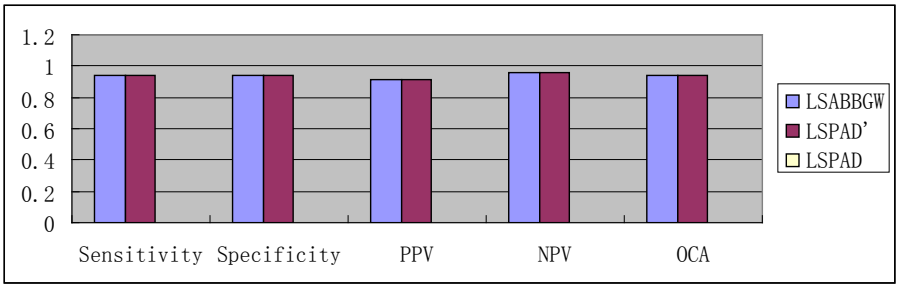


Figure 7. The classification performance comparison on the stem cell matrix dataset using the top 10 features without clustering.

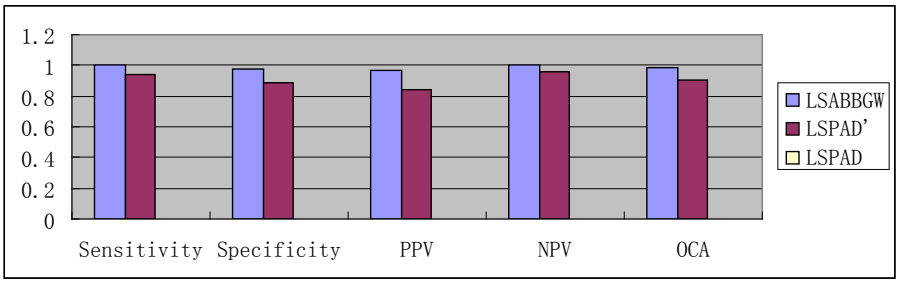


Figure 8. The classification performance comparison on the stem cell matrix dataset using the top 20 Features without clustering.

5. Conclusion

In this article, we proposed a new localized statistical approach to deal with biomarkers selection called localized statistic of abundance distribution based on Gaussian window (LSADBGW). Comparing with the localized statistics of protein abundance distribution (LSPAD), LSADBGW employs the more reasonable local statistical window selection strategy and a more generalized and simpler statistical test method. The classification experimental results prove that our approach perform well than LSPAD. In conclusion, we hope that our LSADBGW method could present useful alternatives in the analysis of the omics data.

REFERENCES

[1] D. J. Oliver, B. Nikolau and E. S. Wurtele, "Functional Genomics: High-Throughput mRNA, Protein, and Metabolite Analyses," Elsevier, 2002, pp. 98-106.

[2] N. Ishii and M. Tomita, "Multi-Omics Data-Driven Systems Biology of *E. coli*," Springer, 2009, p. 41.

[3] S. Smit, H. C. J. Hoefsloot and A. K. Smilde, "Statistical Data Processing in Clinical Proteomics," Elsevier, 2008, pp. 77-88.

[4] H. Shin and M. K. Markey, "A Machine Learning Perspective on the Development of Clinical Decision Support Systems Utilizing Mass Spectra of Blood Samples," Elsevier, 2006, pp. 227-248.

[5] I. Guyon and A. Elisseeff, "An Introduction to Variable

and Feature Selection," MIT Press Cambridge, 2003, pp. 1157-1182.

[6] E. Marchiori, *et al.*, "Feature Selection for Classification with Proteomic Data of Mixed Quality," 2005, pp. 1-7.

[7] H. W. Ressom, *et al.*, "Classification Algorithms for Phenotype Prediction in Genomics and Proteomics," NIH Public Access, p. 691.

[8] M. Dakna, *et al.*, "Technical, Bioinformatical and Statistical Aspects of Liquid Chromatography-Mass Spectrometry (LC-MS) and Capillary Electrophoresis-Mass Spectrometry (CE-MS) Based Clinical Proteomics: A Critical Assessment," Elsevier, 2009, pp. 1250-1258.

[9] Chen, J. J., *et al.*, "Gene Selection with Multiple Ordering Criteria," BioMed Central Ltd., 2007, p. 74.

[10] A. Vlahou, *et al.*, "Development of a Novel Proteomic Approach for the Detection of Transitional Cell Carcinoma of the Bladder in Urine," ASIP, 2001, pp. 1491-1502.

[11] M. J. Campa, *et al.*, "Protein Expression Profiling Identifies Macrophage Migration Inhibitory Factor and Cyclophilin A as Potential Molecular Targets in Non-Small Cell Lung Cancer 1," AACR, 2003, pp. 1652-1656.

[12] J. M. Koomen, *et al.*, "Plasma Protein Profiling for Diagnosis of Pancreatic Cancer Reveals the Presence of Host Response Proteins," AACR, 2005, pp. 1110-1118.

[13] J. M. Koomen, *et al.*, "Diagnostic Protein Discovery Using Proteolytic Peptide Targeting and Identification," John Wiley & Sons, Ltd., Chichester, 2004.

[14] K. R. Kozak, *et al.*, "Identification of Biomarkers for Ovarian Cancer Using Strong Anion-Exchange ProteinChips: Potential Use in Diagnosis and Prognosis," National Acad Sciences, 2003, pp. 12343-12348.

[15] W. Zhu, *et al.*, "Detection of Cancer-Specific Markers Amid Massive Mass Spectral Data," National Acad Sciences, 2003, pp. 14666-14671.

[16] T. C. W. Poon, *et al.*, "Comprehensive Proteomic Profiling Identifies Serum Proteomic Signatures for Detection of Hepatocellular Carcinoma and Its Subtypes," American Association of Clinical Chemistry, 2003, p. 752-760.

[17] A. Valerio, *et al.*, "Serum Protein Profiles of Patients with Pancreatic Cancer and Chronic Pancreatitis: Searching for a Diagnostic Protein Pattern," John Wiley & Sons, Ltd., Chichester, 2001.

[18] M. Wagner, D. Naik and A. Pothen, "Protocols for Disease Classification from Mass Spectrometry Data," WILEY-VCH Verlag Weinheim, 2003.

[19] M. Wagner, *et al.*, "Computational Protein Biomarker Prediction: A Case Study for Prostate Cancer," BioMed Central Ltd., 2004, p. 26.

[20] S. Bhattacharyya, *et al.*, "Diagnosis of Pancreatic Cancer Using Serum Proteomic Profiling," 2004, pp. 674-686.

[21] L. H. Cazares, *et al.*, "Normal, Benign, Preneoplastic, and Malignant Prostate Cells Have Distinct Protein Expression Profiles Resolved by Surface Enhanced Laser Desorption/Ionization Mass Spectrometry 1," AACR, 2002, pp. 2541-2552.

[22] J. M. Sorace and M. Zhan, "A Data Review and Re-Assessment of Ovarian Cancer Serum Proteomic Profiling," BioMed Central Ltd., 2003, p. 24.

[23] T. A. Zhukov, *et al.*, "Discovery of Distinct Protein Profiles Specific for Lung Tumors and Pre-Malignant Lung Lesions by SELDI Mass Spectrometry," 2003, p. 267.

[24] R. X. Li, *et al.*, "Localized-Statistical Quantification of Human Serum Proteome Associated with Type 2 Diabetes," Public Library of Science, 2008.

[25] J. K. Eng, A. L. McCormack and J. R. Yates Iii, "An Approach to Correlate Tandem Mass Spectra Data of Peptides with Amino Acid Sequences in a Protein Database," Elsevier Science Pub. Co., New York, 1994, pp. 976-989.

[26] K. Y. Yeung, *et al.*, "Model-Based Clustering and Data Transformations for Gene Expression Data," Oxford University Press, 2001, pp. 977-987.

[27] Y. I. Moon, B. Rajagopalan and U. Lall, "Estimation of Mutual Information Using Kernel Density Estimators," APS, 1995, pp. 2318-2321.

[28] B. W. Silverman, "Density Estimation for Statistics and Data Analysis," Chapman & Hall/CRC, 1986.

[29] F. J. Müller, *et al.*, "Regulatory Networks Define Phenotypic Classes of Human Stem Cell Lines," Nature Publishing Group, 2008, pp. 401-405.

