British Journal of Mathematics & Computer Science
5(1): 65-81, 2015, Article no.BJMCS.2015.006
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Multi-Resolution Analysis Method for IMS Proteomic Data Biomarker Selection and Classification

Lu Xiong[1] **and Don Hong**[1,2*]

[1] *Computational Science Program, Middle Tennessee State Univ., Murfreesboro, TN 37132, USA.*
[2] *College of Sciences, North China University of Technology, Beijing, China.*

Article Information
DOI: 10.9734/BJMCS/2015/9870
Editor(s): (1) Sheng Zhang, Department of Mathematics, Bohai University, Jinzhou, China.
Reviewers: (1) Manoj Kumar Singh, Centre for Interdisciplinary Mathematical Sciences, Faculty of Science, Banaras Hindu University, India. (2) Anonymous, Central Michigan University, USA.
Complete Peer review History:
http://www.sciencedomain.org/review-history.php?iid=707&id=6&aid=6435

Original Research Article

Received: 04 March 2014; Accepted: 28 April 2014; Published: 09 October 2014

Abstract

Even though imaging mass spectrometry (IMS) technique is evolving rapidly, its data analysis capability lags behind. Especially with the improving of IMS data resolution, faster and more accurate data analysis algorithms are required. To meet such challenges in IMS data analysis, an effective and efficient algorithm for IMS data biomarker selection and classification using multi-resolution (wavelet) analysis method is proposed. We first applied wavelet transform to IMS data de-noising. The idea of wavelet pyramid method for image matching was then applied for biomarker selection, in which Jaccard similarity was used to measure the similarity of wavelet coefficients. Last, the Naive Bayes classifier was used for classification based on feature vectors in terms of wavelet coefficients. Performance of the algorithm was evaluated in real data applications. Experimental results show that this multi-resolution method has advantages of fast computing and accuracy.

Keywords: Proteomics; Biomarker selection; Classification; Imaging Mass Spectrometry; Wavelets.

1 Introduction

Imaging mass spectrometry (IMS) is a technique developed from mass spectrometry to visualize the spatial distribution of moieties such as proteins, peptides, metabolites and lipids ([1], [2]). Currently, IMS is one of the few biochemical technologies able to establish the spatial biochemical composition

**Corresponding author: E-mail: don.hong@mtsu.edu*

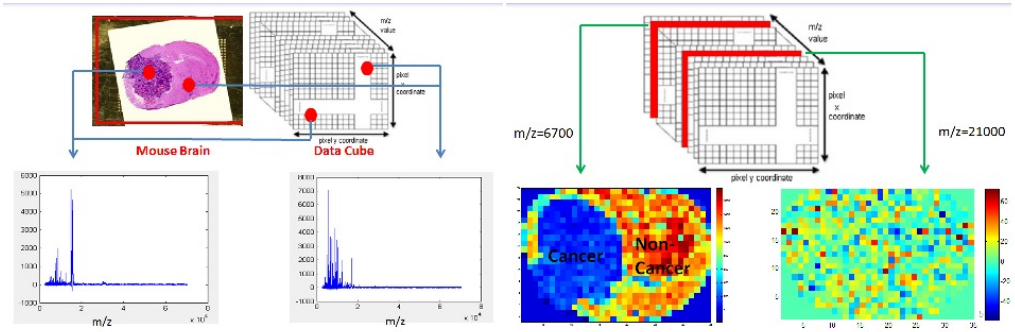

Figure 1: A snap of IMS data. (Left) For a fixed IMS pixel, there is a corresponding mass spectrum (MS). (Right) If an m/z value (mass-to-charge ratio) is fixed, the corresponding MS intensities for all pixels that make up an image shows the spatial intensity of that protein.

of a sample in the full molecular range [3]. It can be used to map biomolecules in biological tissues and has attracted a great deal of attention in the analyses of drug effects, in screening of drugs, and in support for medical diagnoses [4]. However, the development of computational methods for IMS is lagging behind its technological progress [5].

IMS data set can be treated as a hyper-spectral imaging type data cube, see Figure 1. The value at each entry of the IMS data cube shows the abundance of corresponding molecule. For a fixed m/z value (mass-to-charge ratio) in an IMS data cube, the corresponding intensity values make up an image that shows the distribution of that specific biochemical component in sample associated with this m/z value. Also, for a fixed pixel in the image cube, there is a mass spectrum (MS) corresponding to this pixel.

The main tasks for IMS data analysis are biomarker selection and classification. A biomarker is a biological molecule found in blood, other body fluids, or tissues that is a sign of normal or abnormal processes, or of a condition or disease [6]. In IMS data analysis, one usually finds biomarkers in terms of m/z values associated with proteins or peptides. Current popular analysis methods for IMS data include Principle Component Analysis (PCA) ([7], [8]), Support Vector Machine (SVM) [9] and Clustering methods [10]. However, with the development of IMS techniques, the amount and resolution of IMS data has also increased. This requires faster and more accurate data analysis algorithms.

To meet challenges and needs in IMS data analysis, we have developed a mathematical and statistical model using wavelet method for IMS cancer data analysis in biomarker selection and classification. The motivations for introducing wavelet method to IMS data analysis are based on the following. First, the multi-resolution property of wavelets allows us to analyze IMS data on different resolution levels to obtain accurate results with less computation. The low resolution analysis can decrease analysis time because we can represent the whole data set with less wavelet coefficients. Also, over-fitting can be reduced and noises can be lessened at low resolution analysis. The high resolution analysis can improve biomarker selection accuracy by analyzing data without losing detailed information. Wavelet method combines the aforementioned advantages of low and high resolution analysis together. Second, wavelet pyramid idea in image matching [11] can be applied to identify biomarkers from low resolution to high resolution. Note that in cancer IMS studies, biomarkers are identified by comparing cancer IMS data and non-cancer IMS data. This process is similar to image matching. Hence, wavelet method, which is essential in the pyramid imaging matching process, can also be expected to be useful in IMS data analysis. Third, wavelet transformation can reduce the high dimensionality of IMS data. By transforming IMS data to wavelet coefficient space, we can

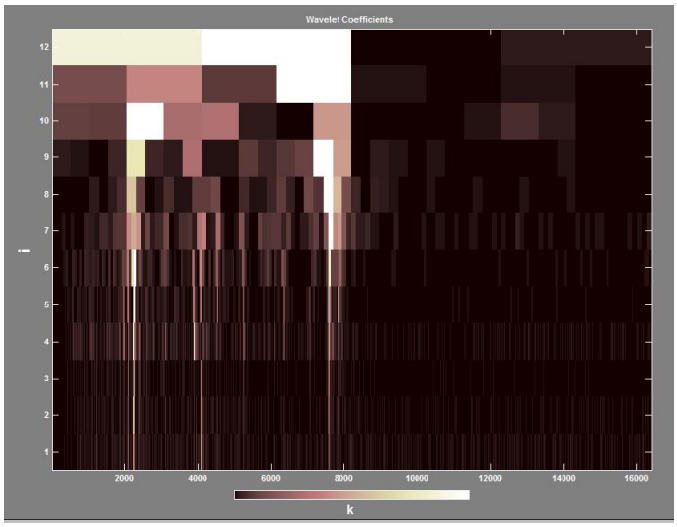

Figure 2: An example of wavelet coefficients of the mass spectrometry for one IMS pixel. Those top coefficients cover wide intervals are low resolution coefficients which describe data on large and rough scale. Those bottom coefficients cover narrow intervals are high resolution coefficients which describe data on small and precise scale.

represent IMS data sparsely at low resolution while still keeping the necessary detail information at high resolution. Last, only few studies have applied wavelet method to IMS data analysis, though there are some work in mass spectrometry (MS) applying wavelet method, for example work in [12]. We would like to apply wavelet method to IMS data to determine if this method has some advantages compares with other current methods. Successful application in MS data analysis would show that the wavelet method can also be promising in dealing with IMS data.

The main contributions of this paper include: combining the advantages of both low resolution and high resolution analysis in IMS data processing to achieve fast and accurate biomarker selection algorithm; providing a new perspective of IMS data by transforming the original IMS data to wavelet coefficient space and can find those patterns not easy to see in original data; introducing probabilistic classification instead of traditional binary classification to obtain not only a classification result but also a confidence level.

The remaining of the paper is organized in the following manner: section 2, we propose a wavelet based de-noise algorithm for IMS data; section 3, a wavelet based IMS biomarker selection algorithm using the idea of pyramid matching is proposed; section 4, we propose an IMS data classification algorithm using feature variables selected from wavelet coefficients combined with Naive Bayes classifier.

2 Wavelet Method for IMS data de-noising

Before we start biomarker selection, we need to pre-process IMS data by data de-noising. It's based on wavelet method. Here is how it works. Figure 2 is an example of wavelet coefficients (discrete Haar wavelet coefficients) for a pixel in IMS data, with false color representation of the coefficient value. The coefficients on the top of Figure 2 are low frequency wavelet coefficients, which describe the data on a large scale and show the outline. The coefficients on the bottom of Figure 2 are high frequency wavelet coefficients, which describe the data on a smaller scale and show the details. In N-

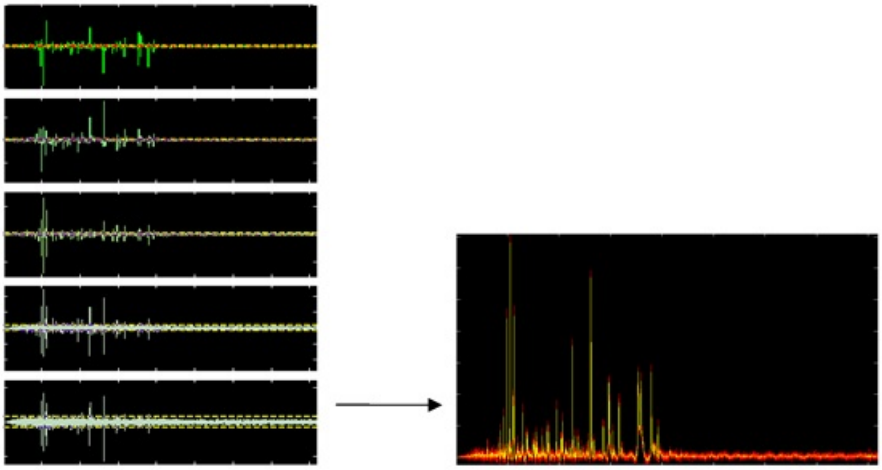

Figure 3: Illustration of IMS data de-noising by using wavelet method. (Left) Apply wavelet transform to mass spectrum. In each resolution level, set a threshold line, the yellow broken lines as shown in left figure. Only keep those large coefficients greater than threshold and set those small coefficients smaller than threshold to be zero. Then apply wavelet inverse transform to the modified coefficients and the result is de-noised data. (Right) The yellow data is de-noised. The red data is original data.

level decomposition, one signal is decomposed into N detailed components and one approximation component. We can de-noise the signal by keeping the large coefficients while setting the small coefficients to be 0 based on a threshold level. By applying this method, we can remove the majority of noises. Here are the basic steps for de-noising,

- **Step 1**: Decompose the signal f. Compute the wavelet decomposition of the signal f from resolution level 1 to N.
- **Step 2**: Threshold detail coefficients. For each level from 1 to N, set the detail coefficients less than threshold to be 0. In illustrative Figure 3, the yellow broken line is the threshold level.
- **Step 3**: Reconstruction of the signal. Compute wavelet reconstruction using the modified coefficients to recover the de-noised signal.

We apply this process to all the original IMS data to obtain the de-noised IMS data. All sequent analysis is based on the de-noised data. See [12] for more details on MS data preprocessing.

3 Biomarker Selection

3.1 Algorithm idea

Biomarkers in IMS cancer studies are proteins whose intensities differ between cancer area tissue and non-cancer area tissue, therefore allowing them to be used as markers to tell the cancer status of the specimen. The biomarker selection problem in IMS data analysis is very similar to the image matching problem. In image matching, people find objects that are similar between images using wavelet pyramid method. Here in IMS data analysis, we find those proteins whose intensities are different between sample data. We just need to define a variable to measure the difference instead

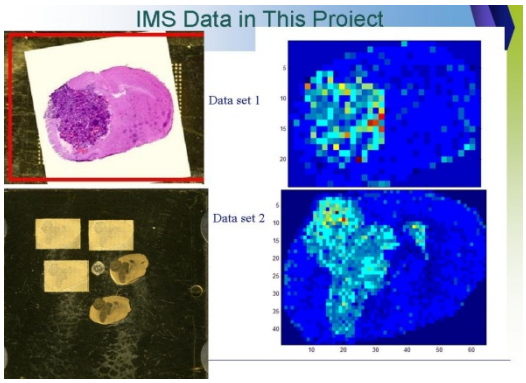

Figure 4: IMS data sets used in this study. (Top left) The brain tissue slice picture where the IMS data set 1 is generated from. (Top right) IMS data set 1 snap for a specific m/z channel. (Down left) The brain tissue slice picture where the IMS data set 2 is generated from. (Down right) IMS data set 2 snap for a specific m/z channel.

of similarity, and biomarker selection problem can be handled in the a similar way as wavelet pyramid method applied in image matching.

The basic idea for image matching based on wavelet pyramid multi-resolution analysis can be briefly described as follows [11].

- **Step 1**: Compare sub-images at the low resolution level.
- **Step 2**: Amplify the matched area and compare images at higher resolution level.
- **Step 3**: Repeat step 2 until to full resolution to find out the matched object in compared two images and purpose of image matching achieved.

We apply this idea to wavelet multi-resolution IMS cancer data analysis to select biomarkers.

- **Step 1**: Compare cancer data and non-cancer data at low resolution level to select the m/z ranges whose intensities are statistically significantly different between cancer and non-cancer data. Those selected m/z ranges can be treated as "suspicious" m/z data ranges because their data difference in statistics may be caused by the existence of cancer biomarkers.
- **Step 2**: Increase the resolution level of those suspicious m/z data ranges to compare them between cancer and non-cancer data at a higher resolution and select those smaller suspicious m/z data sub-ranges with intensity statistically different between two data groups.
- **Step 3**: Repeat step 2 until to full resolution level. Those m/z values selected at full resolution level are the biomarkers we selected from this algorithm.

3.2 Algorithm detail

In this study, we use two IMS data sets as shown in Figure 4. They are generated from the Vanderbilt Mass Spectrometry Research Center using two different mouse brains from same species implanted with the same type of cancer cells. Data set-1 has resolution 24*34, which contains 816 MS pixels. Data set-2 has resolution 64*44, which contains 2816 MS pixels. We use one data set as training data and another as test data. We illustrate this biomarker selection algorithm using the data experiment we did on data set-1. From data set-1 (Figure 4), we select two round IMS data areas with radius of $r = 6$ (6 pixels distance) which are symmetrical to each other by the symmetric line of the mouse brain slice. Because of their symmetrical positions, these two areas contain the very same

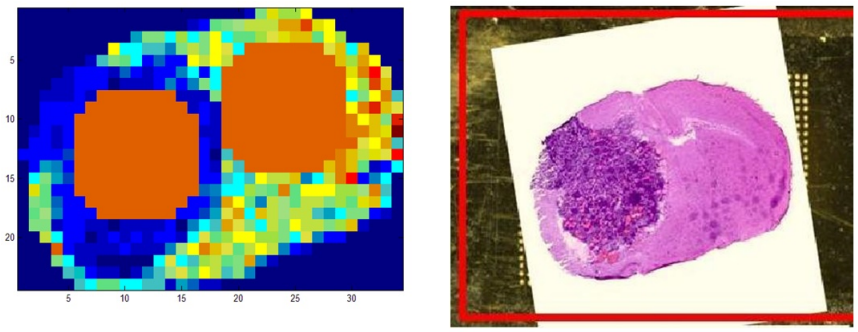

Figure 5: (Left) The left marked round area is the selected cancer pixels and right marked round area is the selected non-cancer pixels; (Right) Slide picture of the mouse brain with a tumor where the data in left was generated.

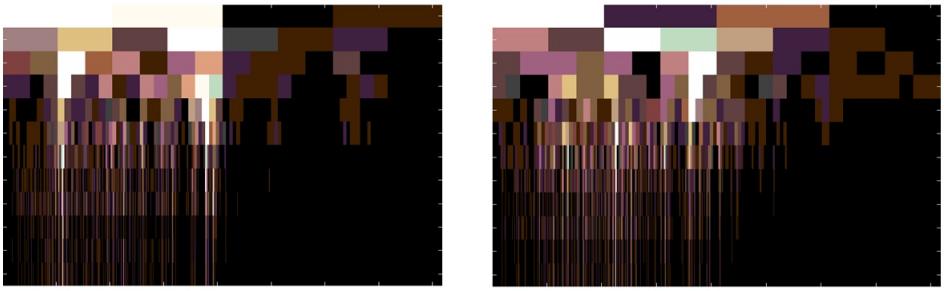

Figure 6: (Left) Wavelet coefficients space for a cancer MS. (Right) Wavelet coefficients space for a non-cancer MS. For each cancer IMS pixel, there is a corresponding wavelet coefficients space like left picture. For each non-cancer IMS pixel, there is a corresponding wavelet coefficients space like right picture. The difference table as shown in Figure 10 is computed by comparing statistical difference of wavelet coefficients from cancer MS group and non-cancer MS group.

biological structure so that we can better emphasize the differentiation of cancer and non-cancer in IMS intensities. The data in these two selected areas are used as training data. Each round area contains 109 IMS pixels, i.e. 109 mass spectra (MS).

For each selected training MS, compute its 12-level discrete wavelet decomposition. Figure 6 shows wavelet coefficients space for a cancer training pixel MS and a non-cancer training pixel MS. Applying wavelet transform to each mass spectrum turns a spectrum data cube into a wavelet coefficients data cube. Originally, each pixel is associated with a mass spectrum, but after transformation, each pixel is associated with a wavelet coefficients vector space. Since the MS intensities of cancer biomarkers vary dramatically from cancer pixels to non-cancer pixels and wavelet coefficient is a description of MS on wavelet space, we can locate biomarkers by measuring the difference between cancer wavelet coefficients and non-cancer wavelet coefficients. The difference of wavelet coefficients can indicate the difference between cancer MS and non-cancer MS at different resolution levels. Analyzing it from low resolution to high resolution, we can quickly locate the biomarkers. This idea was inspired by the wavelet pyramid method in image matching [11].

We measure the difference using a method analogous to Jaccard similarity [13]. It measures difference by measuring how much two groups data are overlapped (Figure 7). Statistically, the more

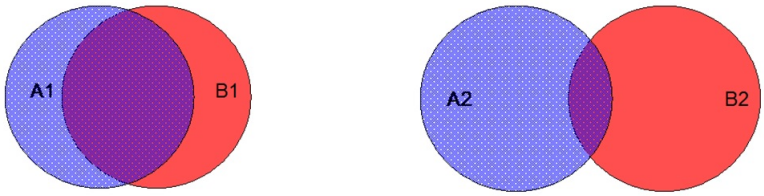

Figure 7: An example of Jaccard similarity. According to definition of Jaccard similarity $J(A,B) = \frac{|A \cap B|}{A \cup B}$, the similarity between A1 and B1 are greater than the difference between A2 and B2. Hence the difference between A1 and B1 are smaller than the difference between A2 and B2.

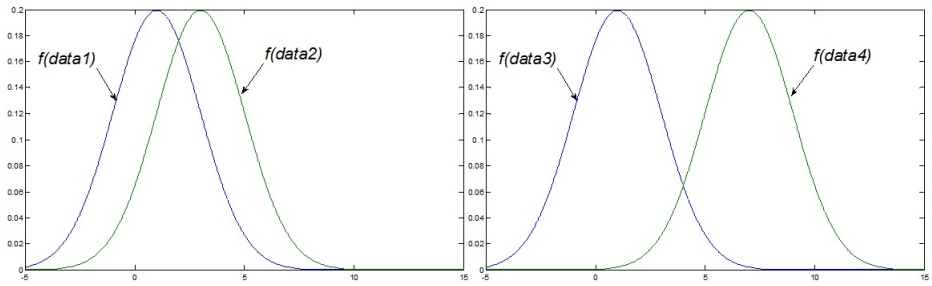

Figure 8: Statistically, according to definition of difference defined in formula (3.1) and (3.2), the difference between $data1$ and $data2$ are smaller than the difference between $data3$ and $data4$, since $data1$ and $data2$ have more overlap values. f is the distribution of data.

two group of data overlap, the more different they are (Figure 8).

The following is the mathematical definition of the difference described above. We denote the set of selected training cancer pixels as S_c, the set of selected training non-cancer pixels as S_n. For a fixed wavelet resolution level $j \in J$ (in our experimental data we define $J = \{1, 2, ..., 12\}$) , a fixed wavelet window position $k \in K$ (in our experimental data we define $K = \{1, 2, ..., 2^{j+2}\}$) and a selected pixel $i \in S_c$ or $i \in S_n$, we denote the corresponding cancer wavelet coefficient as $c^c_{j,k,i}$ and its empirical distribution along the selected training cancer pixels set S_c as $f^c_{j,k}$, and the corresponding non-cancer wavelet coefficients group as $c^n_{j,k,i}$ and its empirical distribution along the selected training non-cancer pixels set S_n as $f^c_{j,k}$. The similarity of wavelet coefficients between cancer data group $\{c^c_{j,k,i}\}_{i \in S_c}$ and non-cancer data group $\{c^n_{j,k,i}\}_{i \in S_c}$ is defined as

$$S_{j,k} = \int_{-\infty}^{+\infty} \min\{f^c_{j,k}(x), f^n_{j,k}(x)\}dx \tag{3.1}$$

The intuitional meaning of $S_{j,k}$ is the overlapping area of the histogram of the two groups to be compared. We can approximately calculate this integral using the histogram of the empirical distribution. Finally, we define the difference between the wavelet coefficients $\{c^c_{j,k,i}\}_{i \in S_c}$ and $\{c^n_{j,k,i}\}_{i \in S_n}$ as:

$$d_{j,k} = 1 - S_{j,k} \tag{3.2}$$

An illustration of $d_{j,k}$ is given in Figure 9.

We define $D = \{d_{j,k}\}_{j \in J, k \in K}$, the difference table that describes the difference of the corresponding wavelet coefficients between cancer group and non-cancer group at different wavelet resolution level

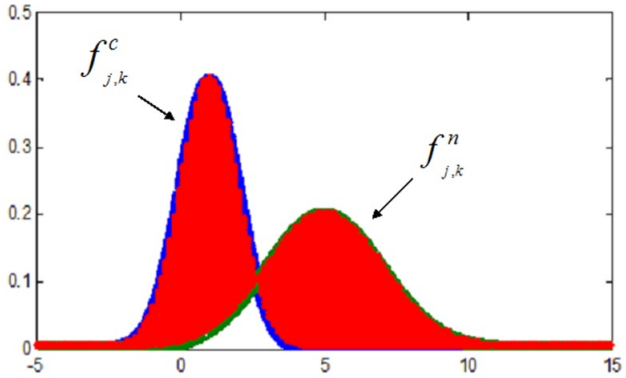

Figure 9: Illustration of the definition of difference $d_{j,k}$, area of red shadow. Since Jaccard distance measure similarity, the difference (un-similarity) should be the complement value of Jaccard distance.

as the false-color map shown in Figure 10. These large difference value areas contain potential biomarkers. "Large difference" means the corresponding $d_{j,k}$ greater than the threshold, i.e., there exists a statistically significant difference here between two groups of data. Similar to the wavelet pyramid method applied in image matching, we take advantage of the multi-resolution property of wavelet analysis to locate the biomarkers from low resolution wavelet coefficients to high resolution wavelet coefficients using the idea we described in section 3.1. We analyze difference table D from lower resolution level $j = 8$ to highest resolution level $j = 12$. From level $j = 8$ to level $j = 12$, if $d_{j,k}$ is greater than the threshold (we set it as 0.85 for the experiment data we use here), that means the contrast between cancer MS and non-cancer MS on the corresponding chemical protein is noticeable at the jth wavelet resolution level as well as the kth wavelet window position. Thus, there is a good chance that biomarkers exist in the corresponding m/z intervals. We then further analyze the wavelet coefficients on the next higher wavelet resolution (i.e. amplify it) level in the same wavelet window position. Otherwise if $d_{j,k}$ is not greater than the threshold, we stop and shift our analysis to the adjacent wavelet window position $k + 1$. We repeat this process until we reach the highest resolution level, level $j = 12$ in the data we used, and determine the specific m/z value whose intensities difference are greater than the threshold. These m/z values selected at highest level $j = 12$ are the m/z values of the biomarkers selected by this algorithm. The threshold can be changed in order to select the corresponding number of biomarkers. Algorithm 1 shows this algorithm's pseudo-code.

Table 1 is the list of the m/z values of the biomarkers selected by this multi-resolution analysis method (MRA) algorithm described above along with the lists of biomarkers selected by some other popular methods for IMS data biomarker selection.

According to the biological study [14], the biomarkers whose $m/z = 6700$ and $m/z = 8380$ are widely confirmed as the key cancer biomarkers for GL26 IMS data sets that we used in this paper. Compared with other methods, the MRA method discovered both biomarkers while including a relatively shorter biomarkers list. Figure 11 is the intensity distributions of these two biomarkers ($m/z = 6702.2$, $m/z = 8374.9$). Its intensity differences between cancer and non-cancer area are significant at this two m/z channels. These are biomarkers that have already been proven in a previous cancer study [15]. Two such biomarkers include cytochrome c oxidase copper chaperone and cytochrome c oxidase subunit 6c. They are related to the growth, division, and expansion of tumor cells. These facts support the results of this MRA algorithm.

Additionally, based on our computing experiment, we determined that the MRA method for IMS data biomarker selection has high algorithm computing speed. We tested the algorithm speed using

Algorithm 1 MRA biomarker selection for IMS data

Require: IMS selected data
Ensure: biomarkers
Compute difference table D ▷ use formula (1), (2)
$lowest\_resolution\_level = 8$ ▷ selected by user
$highest\_resolution\_level = 12$ ▷ selected by user
$desired\_biomarker\_number = 30$ ▷ selected by user
$initial\_threshold = 0.7$ ▷ selected by user
$decrement\_size = 0.01$ ▷ selected by user
$threshold = initial\_threshold$
$biomarkers = []$ ▷ to record biomarkers
while length$(biomarkers) > desired_biomarker_number$ **do**
  for $j = lowest\_resolution\_level \rightarrow highest\_resolution\_level$ **do**
    for $k = 1 \rightarrow 2^{(j+1)}$ **do**
      if $(2^{j+1} * (k-1) + 1, 2^{j+1} * k)$ is in marked interval **then**
        if $D_{j,k} > threshold$ **then**
          Mark $(2^{j+1} * (k-1) + 1, 2^{j+1} * k)$ to be marked interval
        end if
        if $j == highest\_resolution\_level$ **then**
          $biomarkers = [biomarkers, 2^{j+1} * (k-1) + 1]$
        end if
      end if
    end for
  end for
  $threshold = threshold - Decrement\_size$
end while

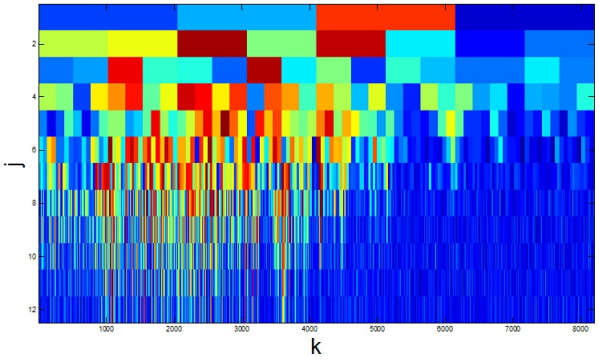

Figure 10: Difference table $D = \{d_{j,k}\}_{j \in J, k \in K}$. The color represents value. It measures the difference between cancer data and non-cancer data. The MRA (multi-resolution analysis) biomarkers selection from low resolution to high resolution is done based on this table.

EN4IMS list	SAM list	EN list	PCA list	MRA list
4664	2791 3434 8337	4476 13562	4934 8567	4599
4667	3010 3764 8366	4664 14327	4936 10257	4607
4670	3056 4011 8380	4670 14336	4937 10259	4757
4812	3734 4076 8395	4812 14343	4938 10261	4759
5446	3800 4271 8492	4884 14781	4939 10263	4762
5753	3920 4538 8672	5425 14786	4960 14969	4767
5754	4206 4566 8945	5429 14805	4962 14971	4770
5756	4341 4665 8982	5446	4963 14974	4892
5757	4605 4676 9327	5753	4964 14976	4895
6165	4734 4899 9343	5754	4966 14979	4903
6702	4767 5106 9531	5756	5439 14981	5438
6706	4921 5120 9602	6165	5441 14983	5446
7799	4936 5428 9619	6702	5442 14986	5449
8019	4964 5444 10238	6706	5444 15603	5714
8024	4981 5707 10267	6794	5445 15606	6244
8384	5001 5753 10466	7799	5446 15608	6248
8386	5024 6166 10662	8019	5448 15611	6312
9344	5170 6186 12434	8024	5449 15613	6702
10172	6225 6251 13560	8028	5451 15616	6705
10261	7706 6310 14525	8384	6571 15618	8375
10263	8420 6574	8386	6572 15620	8400
10265	8603 6700	8495	6574 15623	8403
10267	8709 6719	8524	6575 15625	8572
10282	8747 6780	9344	6577 16780	8978
10366	9062 7099	9553	7749 16782	9332
10374	9736 7118	10172	7751 16785	9613
10825	9956 7297	10261	7752 16787	9616
10949	10167 7315	10263	7792	9624
13562	10952 7338	10267	7794	11632
14336	11388 7357	10282	7795	
14343	11640 7751	10366	7797	
14781	12203 7776	10374	8560	
14786	14865 7795	10811	8562	
14805	14927 8025	10825	8564	
	14978 8107	10949	8566	

Table 1: A comparison of biomarker lists generated by the Multi-Resolution Analysis Method (MRA) and by currently major methods [17] for IMS data analysis. MRA method generates a shorter list while still contains major biomarkers ($m/z = 6702$, $m/z = 8375$).

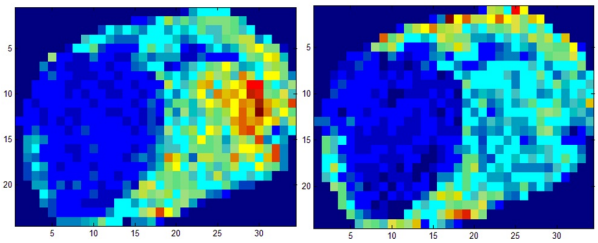

Figure 11: (Left) Intensity distribution for biomarker of $m/z = 6702.2$ selected by MRA method. (Right) Intensity distribution for biomarker of $m/z = 8374.9$ selected by MRA method. These two biomarkers have been confirmed by biology study and also selected out by MRA method.

MATLAB 7.0 installed on a DELL laptop to run EN4IMS proposed by D. Hong and F. Zhang in 2010 [16] and the MRA method discussed in this paper with the same data set (data set-1). Here is the hardware information of the computer used for this test: `Intel(R) Core(TM)2 Duo CPU T7250 @2.00GHZ 778 MHz, 2.00 GB`. The test showed that the CPU time for EN4IMS to select biomarkers is 49.265 seconds. The CPU time for MRA method is only 26.562 seconds. The shorter running time of MRA method comes from the advantage of multi-resolution. In MRA method, we saved computing time by avoiding analyzing every m/z data point one by one. We exclude those m/z intervals whose data difference is not as large as the threshold we set. The amount of m/z data points that still remain at higher resolution levels are much less than the total amount of whole m/z data points. In this way, the amount of data we need to analyze is reduced. Thus, MRA method can achieve high computing efficiency in IMS data analysis.

4 Classification

In this section, we will use the Naive Bayes classifier [18] to do classification on wavelet coefficient space. Bayes classifier is an appropriate tool to deal with IMS data classification problem. It classifies data based on its probabilities in each class and chooses the class with the highest probability to be data's class. Compared with non-probability classification method, Bayes classifier not only tells us a classification result but also the probability to be classified in each class, so we can measure the confidence of results. This advantage is also helpful if we want know the cancer stage or the degree of cancer for each pixel, because more serious degree of cancer corresponds to higher probability to be classified as cancer class in Bayes classifier.

As shown in Figure 12, we use data set-1 as training data and data set-2 as test data for classification study. We train a model from data set-1 and test the trained model using data set-2 to see its performance. Normalization is a necessary step before we start classification, this is because the scale in training data and the scale in test data are different (Figure 13). For normalization purposes, we divide each mass spectrum with its average intensity. After normalization, the scale will be the same in all data sets.

Classification is based on feature variables. We select 10 feature variables from the wavelet coefficients of each pixel's mass spectrum. These feature variables are selected from training data's wavelet coefficients whose values are significantly different between cancer data group and non-cancer data group. We can identify them using the difference table D as shown in Figure 10. Those large entries $d_{j,k}$ in the difference table D correspond to the wavelet coefficients whose difference is large between the cancer group and non-cancer group. Therefore, we chose those wavelet coefficients with large $d_{j,k}$ in D as feature variables. For the data sets used in this study,

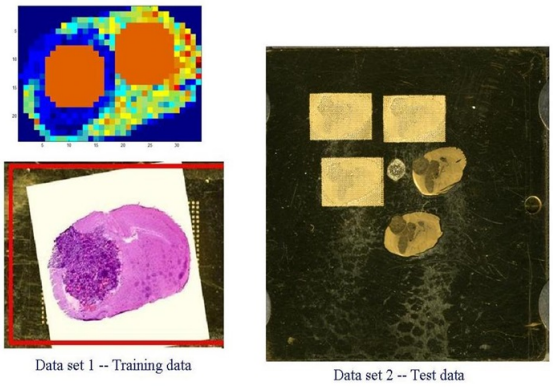

Figure 12: Training data (Left) and test data (Right). The parameters of classification model are trained from training data. The performance of classification model is measured using test data.

we chose wavelet coefficients from level 6 to level 12 whose difference is greater than the difference threshold we set. We can ignore the detail coefficients from level 1 to level 5, since most of the noise exists in high frequency coefficients, if our data contains too much detail, the amount of noise will influence the classification accuracy. With the threshold we set, 10 feature variables are selected from wavelet coefficients space for the mass spectrum of each pixel in cancer and non-cancer training data. These 10 feature variables, as components, form a feature vector, denoted by $\boldsymbol{e}$, corresponding to a pixel MS. The classification process of each pixel is based on its feature vector which made up by its selected feature variables.

In the next step, we use the Naive Bayes classifier to classify the cancer and non-cancer pixels based on pixels feature vector. We denote the probability that an unknown testing pixel i, which has a feature vector $\boldsymbol{X}$, is a cancer pixel as

$$P(i \in C|\boldsymbol{e} = \boldsymbol{X}),$$

where i denotes the testing pixel, C denotes the set of cancer pixels, $\boldsymbol{e}$ denotes the feature vector of this testing pixel, $\boldsymbol{X}$ denotes the value of its feature vector. Similarly, the probability that an unknown testing pixel i, which has a feature vector $\boldsymbol{X}$, is a non-cancer pixel is defined as

$$P(i \in N_c|\boldsymbol{e} = \boldsymbol{X}),$$

where N_c denotes the set of non-cancer pixels. If

$$P(i \in C|\boldsymbol{e} = \boldsymbol{X}) > P(i \in N_c|\boldsymbol{e} = \boldsymbol{X}),$$

the chance of this testing pixel being in the cancer group is greater than its chance in the non-cancer group. If this is the case, then we classify this pixel as a cancer pixel. Otherwise, we classify it as a non-cancer pixel. We can calculate the above conditional probabilities using Bayes formula:

$$P(i \in C|\boldsymbol{e} = \boldsymbol{X}) = \frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X})} \tag{4.1}$$

$$P(i \in N_c|\boldsymbol{e} = \boldsymbol{X}) = \frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)}{P(\boldsymbol{e} = \boldsymbol{X})} \tag{4.2}$$

Then we compare these two probabilities and $P(\boldsymbol{e} = \boldsymbol{X})$ can be canceled, thus leading to:

$$\frac{P(i \in C|\boldsymbol{e} = \boldsymbol{X})}{P(i \in N_c|\boldsymbol{e} = \boldsymbol{X})} = \frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)} \tag{4.3}$$

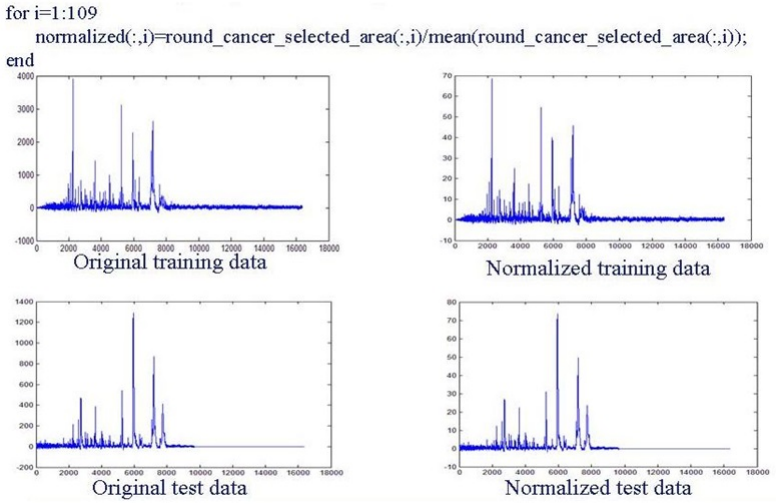

Figure 13: Normalization of Mass Spectrum. Each MS is divided by its average intensity. This is a necessary data preprocessing step before classification. After normalization, the scale of all data sets will be the same.

If $\frac{P(i\in C|\boldsymbol{e}=\boldsymbol{X})}{P(i\in N_c|\boldsymbol{e}=\boldsymbol{X})} > 1$, that means $P(i \in C|\boldsymbol{e} = \boldsymbol{X})$ is greater than $P(i \in N_c|\boldsymbol{e} = \boldsymbol{X})$. We then classify the testing pixel as cancer pixel, since the chance being cancer is larger than the chance being non-cancer. Otherwise, we classify this testing pixel as non-cancer pixel. Therefore, the classification criterion can be defined as:

$$\frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)} > 1 \iff i \in C \tag{4.4}$$

$$\frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)} < 1 \iff i \in N_c \tag{4.5}$$

To calculate values in (4.4), (4.5), we need to determine the likelihood probability $P(\boldsymbol{e} = \boldsymbol{X}|i)$ and find prior probability $P(i)$. Figure 14 shows the distributions of cancer feature variables as well as non-cancer feature variables. They are mostly in normal distributions. Since the feature vector is made up by these 10 feature variables, we can assume that the distribution of feature vector in cancer or in non-cancer is a 10-dimensional normal distribution.

Thus the likelihood $P(\boldsymbol{e} = \boldsymbol{X}|i \in C)$ and $P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)$, which is a probability density, can be calculated by a 10-dimensional normal distribution. The mean value for the cancer data group can be obtained by computing the average value of the feature vectors of all cancer pixels in training data. The standard deviation for the cancer group can be obtained by computing the covariance matrix of the feature vectors of all cancer pixels in the training data. The same idea applies for the non-cancer group. Then, the likelihood for the testing feature vector X can be determined by the remaining of the distributions,

$$(\boldsymbol{e}|i \in C) \backsim N_{10}(\mu_{cancer}, \Sigma_{cancer}) \tag{4.6}$$

$$(\boldsymbol{e}|i \in N_c) \backsim N_{10}(\mu_{noncancer}, \Sigma_{noncancer}) \tag{4.7}$$

where μ, Σ are mean and covariance of the 10-dimensional normal distributions.

To calculate the prior probability $P(i \in C)$ and $P(i \in N_c)$, we count the percentage of each type

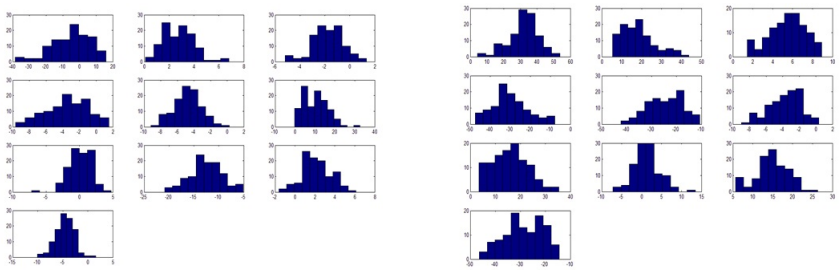

Figure 14: (Left) Distribution of 10 selected feature variables from cancer data group. (Right) Distribution 10 selected feature variables from non-cancer data group. They are mostly approximately normal distributed. Hence it's rational to use 10-dimensional normal distribution to approximate the distribution of feature vector.

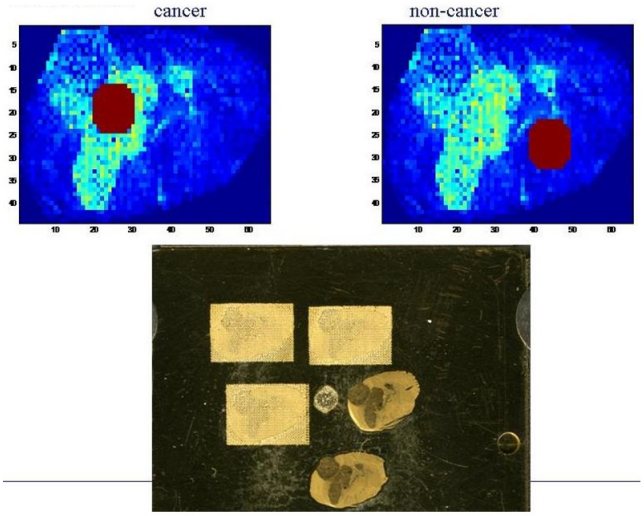

Figure 15: Test data selection. (top left) Cancer test data selection: pixels in the rounded area in this picture are selected cancer pixels; (top right) Non-cancer test data selection: pixels in the rounded area in this picture are selected non-cancer pixels; (down) Photomicrograph of a cresyl violet stained mouse brain section.

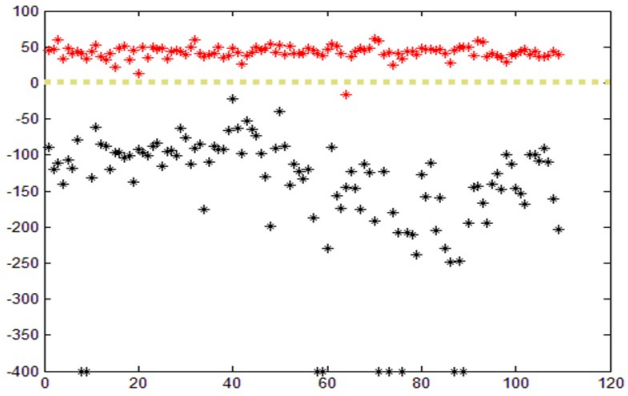

Figure 16: $log_{10}\frac{P(\boldsymbol{e}=\boldsymbol{X}|i\in C)P(i\in C)}{P(\boldsymbol{e}=\boldsymbol{X}|i\in N_c)P(i\in N_c)}$ for test data. According to the classification criteria defined in formula (4.4) and (4.5), 0 is classification boundary (the broken yellow line in this figure). Red points are cancer pixels. Black points are non-cancer pixels. According to classification criteria, those above the yellow broken line should be classified as cancer pixels and those bellow the yellow broken line should be classified as non-cancer pixels.

	Accuracy	Sensitivity	Specificity
PCA+LDA	78.64%	100%	57.27%
PCA+SVM	71.82%	84.56%	59.09%
MRA	99.5%	99.08%	100%

Table 2: Classification algorithm performance of Multi-resolution Analysis Method (MRA) and other popular methods for IMS data analysis, where accuracy represents the rate of correct classification, sensitivity represents the rate that cancer is classified correctly as cancer and specificity represents the rate that non-cancer is classified correctly as non-cancer.

of pixels in training data,

$$P(i \in C) = \frac{|C|}{|C| + |N|} \tag{4.8}$$

$$P(i \in N_c) = \frac{|N|}{|C| + |N|} \tag{4.9}$$

Where $|C|$, $|N|$ are the number of cancer pixels and number of non-cancer pixels in training data respectively. With the above calculations, we can develop a classification model from training data. After we have developed this model using the training data, we test its performance using data set-2 (Figure 15). We select the two rounded marked areas as shown in Figure 15 as test data. Each area contains 109 pixels. The pixels in the left side rounded area are cancer pixels. Those in the right side rounded area are non-cancer pixels.

Figure 16 is the classification result. This graph shows the exponent value of the left side part of formula (4.4) and (4.5) for each pixel. Red points are results for cancer pixels and black points are for non-cancer pixels.

According to the classification criteria defined in formula (4.4) and (4.5), threshold should be 0, since $log_{10}(1) = 0$. Thus, those red points above the threshold and those black points below threshold are classified correctly. According to the result in Figure 16, there is only one pixel in cancer data that is misclassified as non-cancer. Thus, the performance for this classification algorithm is: 99.5% for

accuracy, which represents the rate of correct classification; 99.08% for sensitivity, which represents the rate that cancer is classified correctly as cancer; and 100% for specificity, which represents the rate that non-cancer is classified correctly as non-cancer. Table 2 is a comparison of the performance of Multi-resolution Analysis Method with several other methods.

5 Conclusion

We proposed a multi-resolution analysis (MRA) method for IMS data analysis in biomarker selection and classification. According to data experiment results in table 1 of section 4 and table 2 of section 5, MRA method has advantages in effectiveness and accuracy in biomarker selection and classification comparing with other popular methods. The multi-resolution property of wavelet space saves computating time in finding biomarkers. The data experiment has shown that the CPU computing time of MRA method took only 54% of the computing time using EN4IMS method ([16], [17]).

Though it is challenge to incorporate spatial information for IMS data analysis using MRA method, we will tackle this important problem and report corresponding results in a separate paper.

Acknowledgment

We would like to thank the anonymous referees for their valuable suggestions on the paper. This project was partially supported by Beijing Overseas Talents Program, North China University of Technology. We are also grateful to Vanderbilt Mass Spectrometry Research Center for providing us IMS data in the study.

Competing Interests

The authors declare that no competing interests exist.

References

[1] Rohner T, Staab D, Stoeckli M. MALDI mass spectrometric imaging of biological tissue sections. Mechanisms of Ageing and Development. 2005;126(1):177-185.

[2] McDonnell LA, Heeren R. Imaging mass spectrometry. Mass Spectrometry Reviews. 2007;26(4):606-643.

[3] Trede D, Kobarg JH, Steinhorst K, Alexandrov T Mathematical Methods for Imaging Mass Spectrometry. Cancer Research. 2011;4:5.

[4] Shimma S, Setou M. Review of imaging mass spectrometry. Journal of the Mass Spectrometry Society of Japan. 2005;53:230-238.

[5] Watrous JD, Alexandrov T, Dorrestein PC. The evolving field of imaging mass spectrometry and its impact on future biological research. Journal of Mass Spectrometry. 2011;46(2):209-222.

[6] National Cancer Institute. (2013) NCI Dictionary of Cancer Terms: biomaker. http://www.cancer.gov/dictionary?cdrid=45618. *(Last accessed on 20 December 2013 at 15:36)*.

[7] Van de Plas R, De Moor B, Waelkens E. Imaging Mass Spectrometry Based Exploration of Biochemical Tissue Composition using Peak Intensity Weighted PCA. In Life Science Systems and Applications Workshop, 2007. LISA 2007. IEEE/NIH. IEEE, 2007:209212.

[8] Hanselmann M, Kirchner M, Renard BY, Amstalden ER, Glunde K, Heeren RM, Hamprecht FA. Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis. Analytical Chemistry. 2008;80(24):9649-9658.

[9] Gerhard M, Deininger S, Schleif FM. Statistical classification and visualization of MALDI imaging data. In Computer-Based Medical Systems, 2007. CBMS'07. Twentieth IEEE International Symposium. IEEE, 2007;403-405.

[10] Deininger SO, Ebert MP, Ftterer A, Gerhard M. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. Journal of Proteome Research. 2008;7(12):52305236.

[11] Zavorin I, Le MJ. Use of multiresolution wavelet feature pyramids for automatic registration of multisensor imagery. Image Processing. 2005;14(6):770-782.

[12] Chen S, Hong D, Shyr Y. Wavelet-based procedures for proteomic MS data processing. Computational Statistics and Data Analysis. 2007;52(1);211-220.

[13] Tan P, Steinbach M, Kumar V. Introduction to Data Mining. Addison-Wesley; 2005.

[14] Mayevsky, A. Mitochondrial function and energy metabolism in cancer cells: Past overview and future perspectives. Mitochondrio. 2009;9(3):165179

[15] Matoba S, Kang JG, et al. P53 regulates mitochondrial respiration. Science. 2006;312(5780):1650-1653.

[16] Hong D, Zhang F. Weighted Elastic Net Model for Mass Spectrometry Imaging Processing. Mathematical Modelling of Natural Phenomena. 2010;5(3):808-814.

[17] Zhang F, Hong D. Elastic netbased framework for imaging mass spectrometry data biomarker selection and classification. Statistics in Medicine. 2011;30:753-768.

[18] Domingos P, Pazzani M. On the optimality of the simple Bayesian classier under zero-one loss. Machine Learning. 1997;29(2-3):103130.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=707&id=6&aid=6435