Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme

Weng Howe Chan, Mohd Saberi Mohamad, Safaa Deris, Nazar Zak, Shahreen Kasim, Sigeru Omatu, Juan Manuel Corchado, Hany Al Ashwal

ARTICLE INFO

Article history:
Received 21 March 2016
Received in revised form 3 August 2016
Accepted 3 August 2016

Keywords:
Artificial intelligence
Bioinformatics
Informative genes
Pathway-based microarray analysis
Penalized support vector machine
Weighting scheme
Penalty function

ABSTRACT

Incorporation of pathway knowledge into microarray analysis has brought better biological interpretation of the analysis outcome. However, most pathway data are manually curated without specific biological context. Non-informative genes could be included when the pathway data is used for analysis of context specific data like cancer microarray data. Therefore, efficient identification of informative genes is inevitable. Embedded methods like penalized classifiers have been used for microarray analysis due to their embedded gene selection. This paper proposes an improved penalized support vector machine with absolute t-test weighting scheme to identify informative genes and pathways. Experiments are done on four microarray data sets. The results are compared with previous methods using 10-fold cross validation in terms of accuracy, sensitivity, specificity and F-score. Our method shows consistent improvement over the previous methods and biological validation has been done to elucidate the relation of the selected genes and pathway with the phenotype under study.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The advent of microarray technology has enabled researchers to measure expression of thousands of genes with multiple samples. This has spurred the development of various sophisticated analytical methods to interpret and integrate the enormous data sets produced by high-throughput technology. Moreover, advancements in omics research has made available various kinds of data such as, pathway and network data. Incorporation of pathway knowledge into microarray analysis has brought better understanding and functional insights regarding the biology of the phenotype under study. However, most of the pathway data are curated manually based on public domain databases, domain experts and laboratory study of cultured cells, which are not based on specific biological contexts like lung cancer disease. The use of the pathway data in context specific data like lung cancer microarray data could possess the risk of inclusion of non-informative genes in the analysis, which could affect the analysis outcome. It is known that inclusion of these non-informative genes in classifier construction could lead to poor classification performance. Therefore, methods with effective identification of informative genes in the pathways are needed in order to ensure the efficient utilization of pathway data in aiding the analysis of microarray data.

Generally, there are two types of approaches in pathway-based microarray analysis; the enrichment analysis approach and the machine learning approach. In enrichment analysis, genes are grouped into pathways and scored using statistical tests such as...
Kolmogorov-Smirnov test [8], Fischer’s exact test [9] and gene randomisation [10,11]. Most of the enrichment analysis methods treat all genes equally in the analysis, which is incorrect because some of the genes may have higher relevancy to the phenotype under study which presumably possess better predictive power [12]. On the other hand, machine learning methods often use gene classification in the analysis such as penalized support vector machine [13], random forest [14], partial least square [15] and logistic regression [16]. Furthermore, machine learning methods can be integrated with the gene selection process in order to select informative genes for each pathway before the construction of classifier because not all genes in the pathway contribute to the development of the disease [17,15,18,19]. In order to improve the classification performance, integration of an efficient gene selection method is important to ensure only informative genes are selected for the analysis.

There are three types of gene selection methods commonly known as filter, wrapper and embedded. Embedded methods have been favoured by the researchers owing to their better complexity compared with wrapper methods and interaction with classifier [20]. These embedded methods include random forest, SVM with recursive feature elimination (SVM-RFE) as well as penalized classifier [21]. Among these embedded methods, penalized classifiers have been widely used in bioinformatics research [22]. One of the penalized classifiers is penalized support vector machine (SVM), which is a combination of SVM classifier with penalty functions for simultaneous gene selection. There are several penalized SVM with different penalty functions such as SVM-SCAD, which was proposed by Zhang et al. [23] that embeds smoothly clipped absolute deviation (SCAD) penalty function [24] and L1-SVM [25], which embeds L1 penalty function in standard SVM for feature selection. In a recent published research, a modified penalized SVM with SCAD penalty function, namely gSVM-SCAD has been proposed by Misman et al. [13] and used for pathway-based microarray analysis. gSVM-SCAD was inspired by SVM-SCAD. In contrast to SVM-SCAD, Misman et al. [13] has introduced specific tuning parameters for each pathway to achieve near optimal gene selection. However, performance of the pathway-based methods is often affected by the biological context-free pathway data. Despite the good performance shown by gSVM-SCAD, the gene selection efficiency can still be improved by integrating information derived from the measurements associated with the genes in the pathway. Previous investigation has shown that the use of non-uniform weights calculated from the measurements associated with the genes can improve the identification of informative genes in the pathway [26]. Efficient identification of informative genes in the pathway is crucial especially in the analysis of cancer gene expression data. Therefore, this paper proposes an improved gSVM-SCAD with the integration of absT weighting scheme [26] to improve the efficiency of the identification of informative genes and pathways. The proposed method is referred to as wgSVM-SCAD. As the analysis is done in terms of pathways, the pathway with the most informative genes is expected to produce the best classifier and is identified as the informative pathway related to the phenotype under study.

2. Methods and materials

2.1. gSVM-SCAD

gSVM-SCAD was proposed by Misman et al. [13] for pathway-based microarray analysis. Group-specific tuning parameter, \( \lambda_k \) was introduced for each pathway in order to provide flexibility to SVM-SCAD for maintaining efficient identification of informative genes in every pathway. Given a gene expression data set with \( m \) samples and \( d \) genes \((x_i, y_i), i = 1, \ldots, m, y_i \in \{-1, 1\}\) represents the tissue samples with two classes where \( y_i = -1 \) and \( y_i = 1 \). While \( x_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d \) represents the input vector of expression values of \( d \) genes of \( i\)-th sample. SVM is a classifier which separates the classes of interest by maximising the margin between them using a kernel function. Generally, in gSVM-SCAD, the input variables are classified into corresponding classes by the margin of

\[
\min_{\beta, \lambda} \left[ \sum_{i=1}^{m} \max(0, 1 - y_i \beta^T x_i) + \frac{1}{2} \lambda \sum_{k=1}^{d} \sum_{j=1}^{p} \beta_j^2 \right]
\]

where \( \lambda \) is the penalty parameter, \( \beta \) is the weight vector, \( y \) is the response variable, \( x \) is the input variable, \( \lambda \) is the tuning parameter, and \( d \) is the number of genes.

2.2. The proposed method (wgSVM-SCAD)

Despite the good performance of gSVM-SCAD, the gene selection efficiency can still be improved by integrating information derived from the measurements associated with the genes in the pathway. Meanwhile, the use of non-uniform weighting schemes in pathway analysis has been reported by Ha et al. [26] to improve the performance of the analysis. Therefore, this paper proposes the integration of absT weighting scheme into gSVM-SCAD to improve the performance of the identification of the informative genes and pathways. The weights are calculated based on the measurements associated to the genes, which are more preferable so that they can efficiently represent the differential expression between genes in the data set.

In this paper, the main purpose of the integration of the non-uniform weights is to emphasize the differential expression of the genes and it allows better detection of these differentially expressed genes during the gene selection process using penalized support vector machine. The absT weights are calculated based on two-sample t-test as shown in the equation below [26].

\[
W_j = \frac{|T_j|}{\sum_{i=1}^{m} |T_i|}
\]

In this equation, \( |T_j| \) represents the absolute value of the two-sample t-test calculated for the \( j \)-th gene. The weight of the \( j \)-th gene, \( W_j \) is calculated by dividing \(|T_j|\) with the sum of all \(|T_i|\) for \( d \) genes in the pathway. In this paper, as the weights are calculated
based on the genes in a pathway, the weights indicate how different a gene is expressed between two groups of samples in a pathway. Therefore, the gene that is most differentially expressed in the pathway will obtain the largest weight. The pathway with the most differentially expressed genes will obtain highest classification accuracy. The calculation of the \( \text{absT} \) weights is as illustrated in Fig. 1.

As shown in Fig. 1, for each gene, the weight is calculated by dividing the absolute two-sample \( t \)-test of the gene by the same \( \text{sum} \), which is the sum of the absolute two-sample \( t \)-test of all genes in the pathway. During the calculation process, the \( p \)-value for each gene is recorded. After the calculation of the \( \text{absT} \) weights, an initial filtration is done, which is aimed to filter out genes that do not differentially express between two groups of samples based on the \( p \)-values generated (larger than 0.5) during the two-sample \( t \)-test. Then, the corresponding weights are assigned into the remaining genes and used for further analysis. One weight is given to each of the gene in the pathway. The weight is assigned into the gene expression values of the gene across all the samples as shown in the example in Fig. 2. This produces a weighted gene expression for the pathway.

The weighted gene expression data of the pathway is then used to train an initial linear SVM model and undergoes the penalization based on the genes in a pathway, the weights indicate how different a gene is expressed between two groups of samples in a pathway. Therefore, the gene that is most differentially expressed in the pathway will obtain the largest weight. The pathway with the most differentially expressed genes will obtain highest classification accuracy. The calculation of the \( \text{absT} \) weights is as illustrated in Fig. 1.

As shown in Fig. 1, for each gene, the weight is calculated by dividing the absolute two-sample \( t \)-test of the gene by the same \( \text{sum} \), which is the sum of the absolute two-sample \( t \)-test of all genes in the pathway. During the calculation process, the \( p \)-value for each gene is recorded. After the calculation of the \( \text{absT} \) weights, an initial filtration is done, which is aimed to filter out genes that do not differentially express between two groups of samples based on the \( p \)-values generated (larger than 0.5) during the two-sample \( t \)-test. Then, the corresponding weights are assigned into the remaining genes and used for further analysis. One weight is given to each of the gene in the pathway. The weight is assigned into the gene expression values of the gene across all the samples as shown in the example in Fig. 2. This produces a weighted gene expression for the pathway.

The weighted gene expression data of the pathway is then used to train an initial linear SVM model and undergoes the penalization based on the predefined \( \lambda \). Since the input is the weighted gene expression data, the classification of the proposed method is represented by following equation in contrast to Eq. (1).

\[
\min_{\beta, c} \sum \left[ 1 - y f (Wx) \right] + \text{pen}_\lambda (\beta)
\]

where \( Wx \) represents the weighted input based on the calculated \( \text{absT} \) weights. Similar to the workflow in gSVM-SCAD, the models generated based on different \( \lambda \) in the predefined list are evaluated through GACV and the best fit model is used for performance measurement based on 10-fold CV. The detailed flowchart of the proposed method is shown in Fig. 3.

Compared to gSVM-SCAD by Misman et al. [13], the proposed method integrates non-uniform weights that measure the magnitude of the differential expression of each gene within a pathway based on the concept that the most differentially expressed gene within the pathway have the largest weight in order to improve the efficiency of the identification of informative genes within the pathway. The calculated weights are then integrated into the gene expression data for every gene throughout the samples. Efficient identification of these informative genes improves classification performance, which also enables the identification of informative pathways that are relevant to the phenotype under study. The proposed method is aimed to surmount the limitation caused by the existence of non-informative genes in the pathway data and at the same time provides better biological interpretation regarding the phenotype under study.

2.3. Data sets

This paper uses two types of data, which are gene expression data and pathway data. For gene expression data, four gene expression data sets are used as shown in Table 1. Generally, gene expression data consists of \( m \) samples and \( n \) genes. The data is in the form of matrix where the rows represent the genes while the columns represent the samples. These data sets have been used in pathway-based microarray analysis and can be downloaded at Gene Set Enrichment Analysis (GSEA) website (http://www.broadinstitute.org/gsea/datasets.jsp). According to Table 1, there are two gene expression data sets for lung cancers, Lung Michigan and Lung Boston, which was published by both Beer et al. [28] and Bhattacharjee et al. [29] respectively. While for the gender data set, it consists of transcriptional profiles from male and female lymphoblastoid cell lines. This unpublished data has been used in previous studies in pathway-based microarray analysis [11,13,14]. Lastly, the p53 data set used in this research consists of mutational status of p53 gene in the expression patterns from the NCI-60 cancer cell lines and has been used in several studies in pathway-based analysis [11,13,14].

Meanwhile, for pathway data a total of 480 pathways or gene sets are used similar to the previous works [13,14], which consist of 168 pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) [31] where majority of the pathways were related to metabolism, degradation, biosynthesis and signal processing, and 312 Biocarta pathways [32] where the pathways are mostly related to metabolism and signal transduction. The pathway data are
available for download at MSigDB (url: www.broadinstitute.org/gsea/download.js) [11].

3. Results and discussion

3.1. 10-fold cross validation

The results of wgSVM-SCAD from the 10-fold CV for all the data sets are compared with four embedded methods, which are gSVM-SCAD, L1-SVM, SVM-SCAD, and PathwayRF. L1-SVM is a penalized

Table 1
Summary of gene expression data sets used.

<table>
<thead>
<tr>
<th>Name</th>
<th>Samples</th>
<th>Genes</th>
<th>Class</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Michigan</td>
<td>86</td>
<td>7129</td>
<td>2 (normal/tumour)</td>
<td>Beer et al. [28]</td>
</tr>
<tr>
<td>Lung Boston</td>
<td>62</td>
<td>12,600</td>
<td>2 (normal/tumour)</td>
<td>Bhattacharjee et al. [29]</td>
</tr>
<tr>
<td>Gender</td>
<td>32</td>
<td>22,283</td>
<td>2 (male/female)</td>
<td>Unpublished</td>
</tr>
<tr>
<td>p53</td>
<td>50</td>
<td>12,625</td>
<td>2 (p53+/p53 mutant)</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>

Table 2
Comparisons of average 10-fold CV classification accuracy of top ten pathways with highest accuracy between wgSVM-SCAD and four embedded methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lung Michigan</td>
</tr>
<tr>
<td>wgSVM-SCAD</td>
<td>80.86 (79.66, 80.48)</td>
</tr>
<tr>
<td>gSVM-SCAD</td>
<td>73.77 (—)</td>
</tr>
<tr>
<td>PathwayRF</td>
<td>71.00 (—)</td>
</tr>
<tr>
<td>L1-SVM</td>
<td>55.14 (—)</td>
</tr>
<tr>
<td>SVM-SCAD</td>
<td>53.50 (—)</td>
</tr>
</tbody>
</table>

Note:
- Values in **bold** are highest accuracy.
- Values in the parenthesis are the 95% CI of the accuracy for each method for 10 experiment samples.
- Accuracy values are based on the published results from [13].
SVM classifier using L1 penalty function proposed by Zhu et al. [25]. While PathwayRF is the embedded method based on random forest developed for pathway analysis [14], Table 2 shows the comparison of the performance in terms of average 10-fold CV accuracy of the ten pathways with highest classification accuracy. Comparison of average accuracy for Lung Michigan and gender data sets are based on the results in [13]. Whereas the comparison of average accuracy for Lung Boston and p53 data sets are based on self-experiments. The 95% CI of the accuracy are listed for the results from self-experiments. Based on the data shown in Table 2, wgSVM-SCAD obtains consistent improvements throughout the data sets except in p53 data set, where the improvement of accuracy is not that obvious compared to gSVM-SCAD and PathwayRF.

First, compared with gSVM-SCAD, the better results obtained by the proposed wgSVM-SCAD implies the effectiveness of the introduced abs-T weights and the initial filter based on the p-value of the two-sample t-test in improving the identification of informative genes in the pathway. Meanwhile, in comparison with PathwayRF, which is an embedded method developed for pathway-based microarray analysis, the proposed wgSVM-SCAD obtains 9.86%, 6.67%, 7.00%, 0.62% higher accuracy in Lung Michigan, Lung Boston, Gender and p53 data sets respectively. This is due to the better gene selection behaviour in wgSVM-SCAD that consists of a group-specific parameter for every pathway and in the same time enhanced by the integrated abs-T weighting process. As for L1-SVM, better result from wgSVM-SCAD shows the potency of the SCAD penalty function in providing nearly unbiased estimation for large coefficients compared to LASSO penalty function, which leads to more consistent gene selection [24]. Lastly, compared with SVM-SCAD, flexibility from the group-specific parameter, λk and the assigned abs-T weights have provided better identification of subset of informative genes that are closely related to the phenotype under study. Identification of these informative genes helps to build better classifier and thus better classification accuracy. Furthermore, according to Table 2, the improvement of results from wgSVM-SCAD on Lung Michigan data set is generally higher (for an average difference of 17.51%) in contrast to the results on Lung Boston, Gender and p53 data sets where the corresponding average differences of accuracy are 5.13%, 5.76%, and 3.10% respectively. The main difference between Lung Michigan data and other data sets is the distribution of classes. In Lung Michigan, the sample class distribution is more imbalanced (62 normal samples vs. 24 tumour samples). This also suggests that the proposed wgSVM-SCAD also performs well on data sets with imbalanced classes.

This research aims to develop an improved gSVM-SCAD with better identification of informative genes and pathways. Therefore, in order to analyse and justify the performance of the classifier built, a more comprehensive comparison is done between the proposed wgSVM-SCAD and gSVM-SCAD in terms of sensitivity, specificity and F-score. Both sensitivity and specificity measure how accurate a positive and negative sample is classified respectively. While F-Score is the harmonic measure between sensitivity and precision. A total of 10 runs of full experiment are done for both wgSVM-SCAD and gSVM-SCAD in order to show the consistency of the results of the proposed method. Table 3 shows the comparison of the average sensitivity, specificity and F-Score between wgSVM-SCAD and gSVM-SCAD for all data sets. The complete data and details for the experiments can be found in the supplementary files (Supp. Tables A1 and A4).

For Lung Michigan data, the average difference of sensitivity for the 10 runs is significantly lower compared to gSVM-SCAD at 1.06%. Despite of the low sensitivity, result from the proposed wgSVM-SCAD obtains a significant boost in specificity with an average of 16.28%, which supported by the small p-value as well as the 95% confidence interval (CI). As mentioned previously, the Lung Michigan data set consists of imbalanced class distribution, the result suggests that the introduced abs-T weights have improved the identification of tumour samples because specificity is the measurement of how accurate a negative sample (in this case refers to tumour samples) is being classified. Furthermore, genes discarded during the filtration process could have increase the prediction power of the remaining genes in the pathway. At the same time, the discarded genes might cause the drop in sensitivity, where information from the discarded genes could have better prediction power on normal samples or positive samples. Meanwhile, F-score from the proposed wgSVM-SCAD has shown significant and consistent improvement (average difference of 1.98%) over the previous method. According to the 95% CI, there is 95% chance that improvement in F-Score of the proposed method over the previous method is within 1.60% and 2.36%.

For Lung Boston data set, the proposed wgSVM-SCAD shows consistent and significant improvement in both sensitivity and

Table 3
Average sensitivity, specificity and F-Score from the ten runs of experiments for both wgSVM-SCAD and gSVM-SCAD.

<table>
<thead>
<tr>
<th></th>
<th>Average sensitivity (%)</th>
<th>Difference (%)</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Michigan</td>
<td>wgSVM-SCAD</td>
<td>89.59</td>
<td>90.65</td>
<td>1.71E-02</td>
</tr>
<tr>
<td></td>
<td>gSVM-SCAD</td>
<td>89.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Boston</td>
<td>wgSVM-SCAD</td>
<td>70.82</td>
<td>65.21</td>
<td>5.61</td>
</tr>
<tr>
<td></td>
<td>gSVM-SCAD</td>
<td>65.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>wgSVM-SCAD</td>
<td>93.10</td>
<td>89.40</td>
<td>3.70</td>
</tr>
<tr>
<td></td>
<td>gSVM-SCAD</td>
<td>89.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Average specificity (%)</td>
<td>wgSVM-SCAD</td>
<td>70.95</td>
<td>68.05</td>
</tr>
<tr>
<td></td>
<td>gSVM-SCAD</td>
<td>68.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Michigan</td>
<td>Average F-Score (%)</td>
<td>wgSVM-SCAD</td>
<td>89.40</td>
<td>85.35</td>
</tr>
<tr>
<td></td>
<td>gSVM-SCAD</td>
<td>85.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Michigan</td>
<td></td>
<td>55.90</td>
<td>39.62</td>
<td>16.28</td>
</tr>
<tr>
<td>Lung Boston</td>
<td></td>
<td>69.53</td>
<td>69.03</td>
<td>0.50</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>86.10</td>
<td>84.95</td>
<td>1.15</td>
</tr>
<tr>
<td>p53</td>
<td></td>
<td>84.50</td>
<td>85.35</td>
<td>−0.85</td>
</tr>
<tr>
<td>Lung Michigan</td>
<td></td>
<td>87.26</td>
<td>85.28</td>
<td>1.98</td>
</tr>
<tr>
<td>Lung Boston</td>
<td></td>
<td>71.95</td>
<td>67.93</td>
<td>4.02</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>92.17</td>
<td>89.72</td>
<td>2.45</td>
</tr>
<tr>
<td>p53</td>
<td></td>
<td>74.62</td>
<td>72.78</td>
<td>1.84</td>
</tr>
</tbody>
</table>

Note: p-value < 0.05 for significance.
F-Score based on small p-value obtained and the 95% CI. In sensitivity, the proposed wgSVM-SCAD has brought a significant increase of an average of 5.61% compared to gSVM-SCAD. In contrast to lung Michigan, lung Boston data set is well-balanced in the classes distribution (31 normal samples vs. 31 tumour samples). The increase of the performance is due to the better gene selection of the proposed method due to the integrated absT weighting scheme and the p-value filter. In Lung Boston data set, the efficient gene selection was able to select better genes subsets that have better prediction power on normal samples without compromising too much in the prediction performance on tumour samples. Thus, better classification. Meanwhile, F-score obtained by the proposed wgSVM-SCAD in Lung Boston is steadily higher than gSVM-SCAD in an average of 4.03%. The 95% CI shows that in 95% of the time, the proposed method could obtain 3.50% higher F-Score. This indicates that the classifier constructed by the proposed wgSVM-SCAD performs better than gSVM-SCAD in identifying informative genes in pathways.

For the gender data set, the result from wgSVM-SCAD shows increment of an average of 3.70% and 1.15% compared to the results from gSVM-SCAD for both sensitivity and specificity, respectively. Similar with the case of lung Boston data set, the more efficient gene selection provided by the proposed method has led to a significant and consistent improvement on identification of a certain class without compromising the identification of other class. The performance is depending on the class distribution in the data sets. While the average F-score of wgSVM-SCAD is 2.45% higher than the average F-score of gSVM-SCAD in the gender data set. wgSVM-SCAD achieves significant improvement in both sensitivity and F-Score, which supported with the small p-value obtained and the 95% CI.

For p53 data set, the average sensitivity of wgSVM-SCAD is 2.90% higher than gSVM-SCAD but unfortunately, the specificity of wgSVM-SCAD suffers a drop of 0.85% compared to gSVM-SCAD. This could be due to the initial filtration, which has filtered out the marginal informative genes that have better prediction power on negative samples or p53 mutant samples. Meanwhile, the average F-score of wgSVM-SCAD is 1.84% higher than gSVM-SCAD. However, compared with previous method, the proposed wgSVM-SCAD does not shows significant improvement in p53 data set. According to the statistics, overall performance of the proposed method in p53 data set is on par with the previous method. Overall, this implies that the better classifier is constructed in wgSVM-SCAD due to the efficient identification of informative genes in the pathways through the introduced absT weighting scheme and filtration process.

3.2. Biological validation

In the proposed method, gene selection and classification is performed in each pathway. Therefore, the results consist of the list of pathways ranked based on the average 10-fold CV accuracy. Within each pathway, there is a list of selected informative genes which used to build the classifier and achieve the corresponding accuracy.

As shown in Fig. 4, the identified informative genes from the top five pathways with highest average 10-fold CV accuracy are selected and validated through online biological literature and databases in order to show the biological relevance of these genes to the phenotype under study. Meanwhile, Fig. 5 shows the flow of the process of biological validation. Based on Fig. 5, the identified informative genes and pathways by the proposed method in this research are check through biological literatures and databases like Genecards (url: www.genecards.com) [33] for the related publications in biological research that support the relevance of the genes or pathways to the phenotype of study. Genecards is a multifunctional database that contains numerous types of information regarding a single gene. This process iterates throughout the identified genes for the top five ranked pathways in this research. The validated informative genes or pathways will be further discussed in order to get more biological interpretation related to the phenotype of study.

Table 4 shows the top five pathways identified by the proposed method in Lung Michigan data set. The top ranked pathway is a NFAT signalling pathway, known as NFAT and Hypertrophy of the Heart. Although there is no proof for the direct relationship of this pathway to lung cancer, cardiac involvement in lung cancer is common [34]. Furthermore, expression of the nuclear factor of activated T-cell (NFAT) often involve in cancer cells progression including lung cancer [35]. The proposed wgSVM-SCAD has selected 9 informative genes for the classification and 8 of those genes (CALR, CREBBP, HRA5, MYH2, PI3KCA, RAF1, RPS6KB1) are found related to the development of lung cancer. Among the validated genes, there are several marker genes such as CALR gene, which its expression is associated with tumour pathological grade and previous research has shown that it can be used as a biomarker for lung cancer prediction and diagnosis [36], RAF1 gene, which is associated with tumorigenesis and it is an independent prognostic marker for poor survival rates of lung cancer [37], RPS6KB1 gene, which is a prognostic marker for tumour development in lung cancer [38], and CREBBP gene, where mutation or deletion of this gene contributes to the genesis and the progression of lung cancer cells [39]. The second pathway with highest 10-fold CV accuracy is a Vitamin D pathway known as Control of Gene Expression by Vitamin D Receptor (VDR). The nuclear VDR status has been suggested as a prognostic marker in lung cancer [40]. Investigation also shows that deficiency of VDR occurs in patients with lung cancer [41]. Based on the validation through biological literatures, 8 of the 10 identified informative genes (CREBBP, EP300, VDR, MED1, SMARCC1, SMARCC2, TOP2B, TSC2) are validated genes, there are several marker genes such as CALR

![Fig. 4. Illustration of selection of the top five pathways from the results for biological validation.](image)
found related to the development of lung cancer. The third ranked pathway is known as limonene and pinene degradation pathway. Previous investigations show that pinene has antitumor effect against lung cancer cells [62], while limonene is capable in inhibiting metastatic progression of cancer cells [63]. 9 informative genes have been selected by the proposed wgSVM-SCAD and 5 of those genes (ALDH1A1, ALDH2, CYP24A1, HADHA, NAT6) are validated for their involvement in lung cancer. Among these genes, CYP24A1 is a known independent prognostic marker associated with the survival of patients in lung cancer, where overexpression of CYP24A1 abrogates the antiproliferative effects of 1α,25-dihydroxyvitamin D (3), an active form of vitamin D [55]. The fourth ranked pathway is butanoate metabolism pathway. Butanoate also inhibiting metastatic progression of cancer cells [63]. 9 informative genes have been selected by the proposed wgSVM-SCAD and 5 of those genes (ALDH1A1, ALDH2, CYP24A1, HADHA, NAT6) are validated for their involvement in lung cancer. Among these genes, CYP24A1 is a known independent prognostic marker associated with the survival of patients in lung cancer, where overexpression of CYP24A1 abrogates the antiproliferative effects of 1α,25-dihydroxyvitamin D (3), an active form of vitamin D [55]. The fourth ranked pathway is butanoate metabolism pathway. Butanoate also known as butyrate, which is proven its potential therapeutic implications for diseases such as cancers [64]. Furthermore, investigation has revealed that recognition of alteration in metabolism could have prognostic impact in lung cancer [65]. 9 genes have been selected by the proposed wgSVM-SCAD for classification and 3 of these genes (PRDX6, ALDH2, HADHA) are found related to the lung cancer based on the biological literature. Lastly, the fifth ranked pathway is IL-17 signalling pathway. Previous research has shown that Interleukin 17 (IL-17) is associated with the risk of progressive cancers including lung cancer [66]. In this pathway, the proposed wgSVM-SCAD has selected 11 genes for classification and 6 of these selected genes (CD34, CD3D, CD3E, CD3G, IL6, IL8) are found related to the lung cancer. Among the validated genes, CD34 gene has been identified as potential marker of lung cancer naïve cells [58]. It encodes the protein that plays a role in attachment of stem cells to stromal cells and previous investigation has shown that CD34-positive stromal cells are specific in the stroma of lung adenocarcinomas and may play a supportive role in primary lung cancer [67]. For Lung Boston data set, the top five pathways are shown in Table 5. The top ranked pathway is the polycomb repressive complexes (PRC2) pathway. PRC2 has been frequently implicated in human cancer, acting either as oncogenes or tumour suppressors. High expression of PRC2 components in most of the small cell lung cancer (SCLC) has been reported by Murai et al. [68] and it may play a role in genesis in SCLC as well [69]. Recent investigation also shows that PRC2 is a critical regulator of KRAS-driven non-small cell lung carcinoma progression [70]. 5 genes have been selected by the proposed method for classification and 3 of the selected genes (SUZ12, RBBP7, YY1) are found related to the lung cancer. SUZ12 gene has been reported as an oncogene and a potential diagnostic marker in non-small-cell carcinoma (NSCLC), which associated with the promote of lung tumour growth, migration and invasion [71]. While for the RBBP7 gene, it is found that this gene promotes migration ability of lung cancer cells and
associated with cell growth and it has been identified based on the biological literatures. Expression of estrogen responsive protein efp controls cell cycle and breast tumours growth. The regulation of EIF2 is important for translation initiation and protein synthesis. It has been selected by the proposed method in this research. The proposed method in this research has selected 9 genes for classification and 8 of these genes (CCNB1, CDK4, CDK5, CDK6, SFN, SMURF2, TP53, TRIM25) are identified as lung related. CCNB1 and CDK4 genes are among the potential prognostic marker in NSCLC, where both CCNB1 and CDK4 genes are associated with pathogenesis of lung cancer.

Table 5

<table>
<thead>
<tr>
<th>Pathways</th>
<th>No. of genes in the pathway</th>
<th>No. of selected genes</th>
<th>Selected genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The PRC2 complex sets long-term gene silencing</td>
<td>14</td>
<td>5</td>
<td>SUZ12, RBBP7, YY1, EZH1, RBBP4</td>
</tr>
<tr>
<td>through modification of histone tails</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual signal transduction</td>
<td>24</td>
<td>5</td>
<td>PDE6B, RHO, SAG, SAFB, PDE5A</td>
</tr>
<tr>
<td>Glutamate metabolism</td>
<td>30</td>
<td>12</td>
<td>GCLC, GCLM, GLS, GSS, CAD2, GLS2, GLUD1, GYF1, TMDM1, CMDS, ALDH4A1, ALDH5A1</td>
</tr>
<tr>
<td>Estrogen responsive protein efp controls cell cycle</td>
<td>15</td>
<td>9</td>
<td>CCB1, CKD4, CKD5, SFN, SMURF2, TP53, TRIM25</td>
</tr>
<tr>
<td>breast tumours growth</td>
<td></td>
<td></td>
<td>EIF2AK2, EIF2S1, EIF2S3, RPS20</td>
</tr>
<tr>
<td>Regulation of EIF2</td>
<td>11</td>
<td>6</td>
<td>EIF2AK2, EIF2S1, RPS20, EIF5</td>
</tr>
</tbody>
</table>

Note: The genes in **bold** are genes directly related to lung cancer.

it is a novel biomarker and prognostic marker for distant metastasis in NSCLC [72]. The second ranked pathway is a visual signal transduction pathway and development of lung cancer. The third ranked pathway is glutamate metabolism pathway. Cellular metabolism has been targeted by researchers to improve cancer therapeutics because metabolism in cancer cells is significantly different from normal cells [89]. Glutamate regulates proliferation of neurons and previous research suggests that it is also associated with the motility and the invasive growth of tumour cells in lung cancer [90]. 12 genes have been selected by the proposed wgSVM-SCAD and 4 of these genes (GCLC, GCLM, GLS, GSS) are found related to lung cancers according to the biological literature. The fourth ranked pathway is an estrogen signalling pathway. Estrogen is involved in the biology of NSCLC and inhibition of estrogen synthesis has been shown to prevent tumorigenesis and inhibit the growth of lung tumour [91]. The proposed method in this research has selected 9 genes for classification and 8 of these genes (CCNB1, CDK4, CDK5, CDK6, SFN, SMURF2, TP53, TRIM25) are identified as lung related. CCNB1 and CDK4 genes are among the potential prognostic marker in NSCLC, where both CCNB1 and CDK4 genes are associated with pathogenesis of lung cancer [78,79]. Lastly, the fifth ranked pathway is EIF2 regulation pathway. The protein synthesis initiation factor, EIF2 is important for translation initiation and protein synthesis [87]. 6 genes have been selected by the proposed method and 3 of the selected genes (EIF2AK2, EIF2S1, GSK3B) are validated as related to lung cancer based on the biological literatures. Expression of EIF2AK2 gene is associated with cell growth and it has identified as an independent prognostic variable in NSCLC patients, where low expression of EIF2AK2 leads to aggressive behaviour of lung cancer cells [86], while GSK3B gene is associated with tumorigenesis and aberrant expression of GSK3B gene is known as an independent marker of poor prognosis in NSCLC [88].

Table 6 shows the top five pathways identified by wgSVM-SCAD. The top ranked pathway is GNF_FEMALE_GENES pathway. In this pathway, 1 gene, XIST has been selected by the proposed method. XIST gene is a non-protein coding gene exclusively expressed in female that is involved in the X chromosome silencing in female cells and allow X chromosome equilibration with males [92]. A recent study also found that overexpression of XIST is involved in psychiatric disorders in females [93]. The second ranked pathway is TESTIS GENES_FROM_XHX_AND_NETAFFX pathway. Similar to the top ranked pathway, 1 gene has been selected by the proposed method, which is RPS4Y1 and it is found differentially expressed between male and female [94]. RPS4Y1 gene is located in the Y chromosome in males, which encodes for ribosomal protein S4 [95]. While the third ranked pathway is the XINACT pathway, which is an inactivation process of X chromosome to achieve equal dosage of X chromosome genes in males and females [96], 2 genes have been selected by the proposed method and both (DDX3X and RPS4X) are validated through biological literature. The fourth ranked pathway is RAP_DOWN, which is related to the rapamycin down regulation. Previous studies found that rapamycin can extend the lifespan in mice, especially for female mice [97,98]. 3 genes have been selected by the proposed method and 2 of these (DDX3X and HDHD1A) are found to be related in the differentiation between genders. Lastly, the fifth ranked pathway is INSULIN_2F_UP pathway. A previous study has shown the role of insulin in cell proliferation [99]. 3 genes have been selected by the proposed method for classification and 1 gene is validated through the biological literature, which is HDHD1A.

For p53 data set, the top five pathways identified by the proposed method is shown in Table 7. The top ranked pathway is known to influence RAS and RHO proteins on G1 to S transition pathway. P53 is a well-known tumour suppressor protein. Mutation or loss of p53 have been a critical event in many human cancers. Previous investigation has shown that mutation of p53 if coupled with the activated RAS can induces RhoA activity and promotes cancer progression [100]. 10 genes have been selected by wgSVM-SCAD for classification and 9 of the genes (AKT1, RHOA, CDK4, CDKN1A, E2F1, IKBKB, RAC1, RELA, TFF1) are found related to the p53 mutation in the cancer cell lines. The second ranked pathway, is known as Cell Cycle: G1/S Check Point pathway. Previous studies have reported the involvement of p53 in regulation of G1 checkpoint to control cell proliferation by triggering apoptosis in normal cells as well as cancer cells [101,102]. Thus, mutation in p53 leads to the differential expression of this pathway. In

Table 5

<table>
<thead>
<tr>
<th>Pathways</th>
<th>No. of genes in the pathway</th>
<th>No. of selected genes</th>
<th>Selected genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The PRC2 complex sets long-term gene silencing</td>
<td>14</td>
<td>5</td>
<td>SUZ12, RBBP7, YY1, EZH1, RBBP4</td>
</tr>
<tr>
<td>through modification of histone tails</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual signal transduction</td>
<td>24</td>
<td>5</td>
<td>PDE6B, RHO, SAG, SAFB, PDE5A</td>
</tr>
<tr>
<td>Glutamate metabolism</td>
<td>30</td>
<td>12</td>
<td>GCLC, GCLM, GLS, GSS, CAD2, GLS2, GLUD1, GYF1, TMDM1, CMDS, ALDH4A1, ALDH5A1</td>
</tr>
<tr>
<td>Estrogen responsive protein efp controls cell cycle</td>
<td>15</td>
<td>9</td>
<td>CCB1, CKD4, CKD5, SFN, SMURF2, TP53, TRIM25</td>
</tr>
<tr>
<td>breast tumours growth</td>
<td></td>
<td></td>
<td>EIF2AK2, EIF2S1, EIF2S3, RPS20</td>
</tr>
<tr>
<td>Regulation of EIF2</td>
<td>11</td>
<td>6</td>
<td>EIF2AK2, EIF2S1, RPS20, EIF5</td>
</tr>
</tbody>
</table>

Note: The genes in **bold** are genes directly related to lung cancer.

Table 6

<table>
<thead>
<tr>
<th>Pathways</th>
<th>No. of genes in pathway</th>
<th>No. of selected genes</th>
<th>Selected genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNF_FEMALE_GENES</td>
<td>116</td>
<td>1</td>
<td>XIST [93]</td>
</tr>
<tr>
<td>TESTIS_GENES_FROM_XHX_AND_NETAFFX</td>
<td>111</td>
<td>1</td>
<td>RPS4Y1 [94]</td>
</tr>
<tr>
<td>XINACT</td>
<td>34</td>
<td>2</td>
<td>DDX3X, RPS4X [94]</td>
</tr>
<tr>
<td>RAP_DOWN</td>
<td>434</td>
<td>3</td>
<td>HDHD1A, DDX3X [94], RTN4</td>
</tr>
<tr>
<td>INSULIN_2F_UP</td>
<td>405</td>
<td>3</td>
<td>HDHD1A [94], RPS20, RPL41</td>
</tr>
</tbody>
</table>
for some of the orphan receptors have been identified in the lipid metabolism. Furthermore, several lipid ligands have been shown to play a role in lipid metabolism [132], therefore, mutation on P53 could lead to deregulation of the apoptotic pathway.

The Nuclear Receptors in Lipid Metabolism and Toxicity pathway is the system where the proposed method has selected 8 genes that can participate in lipid metabolism, storage, elimination as well as the activation of procarcinogens [133]. The proposed method has selected 8 genes in this pathway for classification, 7 of which (BAX, BCL2, CASP8, RELA, TNFRSF10B) are found related to the p53 mutation in cancer cell lines. The fourth ranked pathway is the Nuclear Receptors in Lipid Metabolism and Toxicity pathway, and the proposed method has selected 8 genes in this pathway for classification, 7 of which (BAX, BCL2, CASP8, RELA, TNFRSF10B) are found related to the p53 mutation in cancer cell lines.

<table>
<thead>
<tr>
<th>Pathways</th>
<th>No. of genes in the pathway</th>
<th>No. of selected genes</th>
<th>Selected genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influences of RAS and RSH proteins on G1 to S Transition</td>
<td>28</td>
<td>10</td>
<td>AKTI [103], RHOA [104], CDK4 [105], CDKN1A [106], EZF1 [107], IKBKB [108], RAC1 [109], RELA [110], TFDP1 [111], PAK1 ABL1 [112], ATR [113], CDK2 [114], CDKN1A [106], CDKN2A [115], CDKN2B [110], DHFR [117], EZF1 [107], HADAC1 [118], SMAD3, TGFBI, BAX [119], BCL2 [120], CASP9 [121], RELA [110], TNFRSF10B [122], DFFA, IL1R1 PPARC [123], RARB [124], VDR [125], CYPIA2, CYPIB7P, CYP2E1, CYP2C9, SFTPBI, MRE11A, ATM [126], ATR [113], BRCA2 [127], CHEK1 [128], NBN [129], RAD50 [130], BRDT, RAD5, FANC G</td>
</tr>
<tr>
<td>Cell cycle: G1/S check point</td>
<td>26</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Apoptosis</td>
<td>81</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Nuclear receptors in lipid metabolism and toxicity</td>
<td>53</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Role of BRCA1 and BRCA2 and ATR in cancer susceptibility</td>
<td>22</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Note: The genes in **bold** are genes related to p53 mutations in cancer cell lines.

4. Discussion

The proposed wgSVM-SCAD is aimed to provide effective gene selection in order to surmount the performance limitation caused by the existence of non-informative genes in the pathways. Based on the 10-fold CV performance shown in the Table 2, the proposed method obtains improved classification accuracy compared to gSVM-SCAD as well as previous studies throughout all the data sets. Furthermore, performance of both wgSVM-SCAD and gSVM-SCAD have been further evaluated in terms of specificity and sensitivity as shown in Tables 3 and 4. The performance of wgSVM-SCAD is consistently improved compared to the results of gSVM-SCAD throughout all the data sets. This indicates that the gene selection efficiency of the proposed method has been increased because the use of informative genes in classifier construction leads to better classification performance. This shows that the introduced abdT weights for every gene in the pathway can improve the gene selection efficiency by emphasizing the differential expression of the genes in the pathway. Moreover, the initial filtration step also helps to filter out non-informative genes based on the p-value. While the remaining weighted genes become better candidates in the selection process by the SCAD penalty function. Besides, notable improvement in the specificity in Lung Michigan data set implies that the introduced abdT weights and initial filtration is able to improve the classification performance on data set with imbalanced class distribution (62 normal samples and 24 tumour samples). Thus, the proposed method, wgSVM-SCAD is a better classifier compared to gSVM-SCAD, which is further supported by the improved F-scores shown in Table 5.

As for biological validation, top five pathways with highest classification accuracy have been validated and shown in previous section. It’s worth mentioning that there are several genes that have been selected more than once in different pathways and even different data sets. These genes have been validated through this pathway, the proposed method has selected 11 genes for classification and 9 of these genes (ABL1, ATR, CDK2, CDKN1A, CDKN2A, CDKN2B, DHFR, EZF1, HADAC1) are identified as closely related to the p53 mutation in cancer cell lines. The third ranked pathway is the apoptosis pathway, P53 has been known for its influence in cell cycle control and apoptosis. Mutation of p53 or loss of p53 could deregulate apoptosis and promotes tumour cell migration and invasion [131]. 7 genes have been selected in this pathway by wgSVM-SCAD for cancer classification. 5 of the selected genes (BAX, BCL2, CASP8, RELA, TNFRSF10B) are found related to the p53 mutation in cancer cell lines. The fourth ranked pathway is the Nuclear Receptors in Lipid Metabolism and Toxicity pathway, and the proposed method has selected 8 genes in this pathway for classification, 7 of which (BAX, BCL2, CASP8, RELA, TNFRSF10B) are found related to the p53 mutation in cancer cell lines. Lastly, the fifth ranked pathway is the BRCA related pathway known as Role of BRCA1 and BRCA2 and ATR in cancer susceptibility. Similar to p53, BRCA1 gene is a tumor suppressor gene involved in deoxyribonucleic acid (DNA) damage response pathway and mutation of BRCA1 gene is common in a variety of cancers. A previous study has reported that expression of BRCA1 is controlled by the presence of p53 [134]. Moreover, it was reported that p53 deficiency is highly cooperative with both BRCA1 and BRCA2 in promoting tumorigenesis [135]. 10 genes have been selected from this pathway by the proposed method for classification and 7 of these genes (MRE11A, ATM, ATR, BRCA2, CHEK1, NBN, RAD50) are found related to the p53 mutation in cancer cell lines.

![Fig. 6. Overlapped genes identified in the top five pathways in Lung Michigan data set.](Image)
biological literature. This indicates that some of the selected genes could be the potential driver genes between two or more processes involved in the development of cancers. Fig. 6 shows the genes that have been identified more than once in the top five pathways in lung Michigan data set. As discussed in previous section, the NFAT pathway and Vitamin D Receptor pathway are associated with the cancer cells progression [40,41,35,34]. While CREBBP gene has been reported for its role in development of lung cancer, where mutation or deletion of this gene has resulted the genesis and progression of lung cancer cells [39]. Meanwhile, HADHA and ALDH2 also have been selected in both limonene and pinene degradation pathway and butanoate metabolism pathway. These two pathways are associated with antitumor effect and the therapeutic implications in lung cancer.

Meanwhile, Fig. 7 shows the overlapped genes that have been selected by the proposed method in different pathways. HDHD1A gene has been selected in both XINACT and RAP_DOWN pathway while DDX3X gene has been selected in both RAP_DOWN and INSULIN_2F_UP pathway. Higher expression of both HDHD1A and DDX3X gene are found in female compared to male according to the previous investigations [94,136]. Recent investigation has shown the role of X inactivation and rapamycin downregulation pathway in gender-specific functions [137,96].

For p53 data set, several genes have been selected in more than one pathway as shown in Fig. 8. RELA gene, where its function is associated with cancer cells progression, has been identified in both apoptosis pathway and RAS and RHO pathway. While CDKN1A and E2F1 have been selected in both RAS and RHO pathway and G1/S check point pathway, which are closely related to the cell cycle control. CDKN1A is involved in the G1 progression control, which associated negative regulation of p53 tumour suppressor [106] while E2F1 plays an important role in making vital cellular decision to regulate the cell cycle progression and defect on the E2F1 and P53 is associated with tumour development [107]. Meanwhile, ATR gene has been selected in both G1/S check point pathway and the BRCA pathway. ATR plays an important role in regulation of cancer cells proliferation.

The weighting scheme used in this paper allows more efficient detection of informative genes and pathways that are differentially expressed between two condition. With the rise of the next-generation sequencing such as RNA-seq, which has shown strong potential to replace microarrays for whole-genome transcriptome profiling. This is because RNA-seq offers better specificity and sensitivity in examining transcriptome fine structure. In this case, the weighting scheme could be beneficial in improving the detection of differential expressed transcript between different conditions. Meanwhile, in comparison with some of the widely used pathway analysis software suite, such as Ingenuity Pathway analysis and GeneSpring, these software suites offer very comprehensive analysis capabilities such as comparative analysis,

\[ \text{Apoptosis} \quad \text{Cell Cycle: G1/S Check Point} \]

\[ \text{RELA} \quad \text{CDKN1A} \quad \text{E2F1} \quad \text{ATR} \]

\[ \text{Influence of RAS and RHO proteins on G1 to S Transition} \quad \text{Role of BRCA1 and BRCA2 and ATR in cancer susceptibility} \]

Fig. 8. Overlapped genes identified in the top five pathways in p53 data set.

5. Conclusion

The proposed wgSVM-SCAD has shown overall improvements over gSVM-SCAD in terms of classification accuracy, sensitivity, specificity and F-Score. Integration of the absT weights improves the performance of the selection, allowing better identification of the informative genes in the pathways related to the phenotype under study and leads to better classification performance. This provides a potential solution to effectively identify informative genes related to the phenotype under study from pathway data that are often not based on specific biological contexts. Furthermore, biological validation of the top five pathways shows that our proposed method is able to identify some of the marker genes as well as genes that play important roles in the development of the phenotype under study. Despite the overall improvements of the proposed wgSVM-SCAD over gSVM-SCAD, the introduction of absT weights and the initial filtration have led to the slight performance drop on sensitivity in lung Michigan data set, and specificity in lung Boston, gender and p53 data sets. This could be due to the sub-optimal choice of parameter $\lambda$ in SCAD because SCAD penalty is parametric. Therefore, its performance of gene selection is relying on the choice of $\lambda$. When the $\lambda$ is too big, underfitting of data may occur and produces a very sparse classifier. When the $\lambda$ is too small, there is risk of overfitting the data and produces a less sparse classifier. As the proposed method uses a grid search of predefined set of $\lambda$ for SCAD, improvement could be done by integrating more robust penalty with better $\lambda$ selection using stochastic search such as metaheuristics.

Conflict of interest

None declared.

Acknowledgement

This research is supported by Malaysian Ministry of Education through several Fundamental Research Grant Schemes (Grant
Appendix A
Supporting information

Supplemental data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compiobiomedi.2016.08.004.

References


J. Ma, Y. Xue, W. Cui, Y. Li, Q. Zhao, W. Ye, J. Zheng, Y. Cheng, Y. Ma, S. Li, T. Han, L. Miao, L. Yao, J. Zhang, W. Liu, Ras homolog gene homolog, member A promotes p53 degradation and vascular endothelial growth factor-dependent angiogenesis through an interaction with murine double minute 2 under hypoxic conditions, Cancer 118 (17) (2012) 4010–4016.


Nazar Zaki is an Associate Professor and coordinator of two tracks namely Intelligent Systems and Software Development. His research interest is in the field of bioinformatics, data mining and machine learning. He mainly focuses on developing an intelligent data mining algorithms to solve specific biological problems, such as protein function/structure prediction, protein interaction network analysis and protein complex detection.

Shahreen Kasim is a Senior Lecturer at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. She received the B.Sc., M.Sc. and Ph.D. degrees in Computer Science from the Universiti Teknologi Malaysia in 2003, 2005 and 2011, respectively. Her research interests focus on gene function prediction, clustering algorithm and computational biology.

Sigeru Omatu has been a Professor at Osaka Institute of Technology since 2010. He received his Ph.D. in Electronic Engineering from Osaka Prefecture University. He was a Professor at the University of Tokushima in 1988 and Osaka Prefecture University in 1995. His research area is intelligent signal processing based on neural networks.

Juan Manuel Corchado is a full Professor with chair at the University of Salamanca. He holds a Ph.D. in Computer Sciences from the University of Salamanca and a Ph.D. in Artificial Intelligence from the University of the West of Scotland. His interests are in areas of information fusion, intelligent distributed systems, ambient intelligence and sensors.

Hany Al Ashwal joined UAEU in 2015 as Assistant Professor in the College of Information Technology. His research interests focus on gene expression data analysis and bioinformatics applications in Biomedical and Pharmaceutical Sciences. Research interests also include artificial intelligence, machine learning, genome wide analysis, and computational biology.