<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>A New Method for Preliminary Identification of Gene Regulatory Networks from Gene Microarray Cancer Data Using Ridge Partial Least Squares with Recursive Feature Elimination and Novel Brier and Occurrence Probability Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Chan, SC; Wu, HC; Tsui, KM</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2012, v. 42 n. 6, p. 1514-1528</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2012</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/164063">http://hdl.handle.net/10722/164063</a></td>
</tr>
</tbody>
</table>

S. C. Chan, Member, IEEE, H. C. Wu, Student Member, IEEE, and K. M. Tsui

Abstract—This paper proposes a new method for preliminary identification of gene regulatory networks (GRNs) from gene microarray cancer data based on ridge partial least squares (RPLS) with recursive feature elimination (RFE) and novel Brier and occurrence probability measures. It facilitates the preliminary identification of meaningful pathways and genes for a specific disease, rather than focusing on selecting a small set of genes for classification purposes as in conventional studies. First, RFE and a novel Brier error measure are incorporated in RPLS to reduce the estimation variance using a two-nested cross validation (CV) approach. Second, novel Brier and occurrence probability-based measures are employed in ranking genes across different CV subsamples. It helps to detect different GRNs from correlated genes which consistently appear in the ranking lists. Therefore, unlike most conventional approaches that emphasize the best classification using a small gene set, the proposed approach is able to simultaneously offer good classification accuracy and identify a more comprehensive set of genes and their associated GRNs. Experimental results on the analysis of three publicly available cancer data sets, namely leukemia, colon, and prostate, show that very stable gene sets from different but relevant GRNs can be identified, and most of them are found to be of biological significance according to previous findings in biological experiments. These suggest that the proposed approach may serve as a useful tool for preliminary identification of genes and their associated GRNs of a particular disease for further biological studies using microarray or similar data.

Index Terms—Cancer classification, gene regulatory network (GRN), logistic regression (LR), microarray data, novel Brier and occurrence probability (OP) measures, ridge partial least squares (RPLS) regression.

I. INTRODUCTION

The study of gene regulatory networks (GRNs) has received considerable attention in systems biology because of its potential in explaining various biological processes. Such information could be very useful to clinical applications, such as the prediction of cancer and other diseases, etc. Many methods have been proposed for stability [1], modeling [2]–[6], and inference [7], [8] of GRNs.

In this paper, we focus on the detection of statistically more influential genes and utilize them for preliminary identification of GRNs associated with specific cancer using gene microarray data. A gene microarray sample consists of an observation (e.g., the subject with/without cancer for binary classification) and a set of gene expression intensities. Due to the high cost of obtaining the gene microarray data and the limited number of patients, such microarray data usually contain small number of samples but a large number of gene expressions. Therefore, even the identification of a few marker genes presents great challenges due to the curse of dimensionality and limited samples.

Conventionally, gene identification is considered as a feature selection problem, and two well-known types of feature selection algorithms are the filter-based and wrapper-based algorithms. Typical filter-based algorithms rely on measures such as the signal-to-noise ratio (SNR) [9], t-statistics (t-test) [10], Fisher’s criterion (FC) [11], point biserial correlation coefficients (PC) [12], and so on to select the important genes, and discard those with lower scores. One of the major drawbacks of the filter-based algorithms is that the gene expressions are assumed to be independent, but in practice, they are usually correlated due to their pathway dependencies [13]. On the other hand, many wrapper-based algorithms such as [14]–[17] have been proposed for gene selection aiming at better classification performance.

While most algorithms focus on identifying a small gene set for the best classification performance, we take a different approach in this paper to study the detection of a larger set of cancer-related genes so as to preliminarily identify the GRNs associated with the cancer. One of the major challenges on identifying the GRNs is to tackle the pathway dependencies among the genes. A master control gene, which is an active gene at the top of a control hierarchy, usually shares a common pathway with some amplifier genes that serve to amplify the signal on the same GRN. When the master control gene is expressed, it triggers the amplifier genes to amplify the signal on the GRN, which might further trigger a cascade of other

Manuscript received December 1, 2010; revised July 7, 2011 and March 7, 2012; accepted March 29, 2012. Date of current version October 12, 2012. This paper was recommended by Associate Editor R. van Paassen.

The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong (e-mail: scchan@eee.hku.hk; andrewhwwu@eee.hku.hk; kmtsui@eee.hku.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2012.2199302
GRNs and turn on a huge number of other genes [18]. An illustrative example in [19] suggests that a single signal from a master control gene can trigger the out-of-place eye production in the fruit fly Drosophila. Therefore, the expression of the genes on the same GRN are correlated, and a subtle change in the expression of the master control gene may have strong physiological consequence, while its change of expression may be hidden by other related amplifier genes on the same GRN [13].

Biological and statistical variability is another challenge for the identification of GRNs and cancer-related genes. Conventionally, there are two approaches in ranking genes in regression techniques. The simplest one is to invoke the algorithm on the whole data set, i.e., use the whole data set as training, and select the subsets with the best performance. However, due to small number of samples, the estimation is subject to large variance though the classification accuracy on the samples themselves may be high, i.e., it lacks generalization ability due to overfitting. Another more reasonable approach reported in [20] is to employ cross validation (CV) to evaluate the performance of the classifier where the data are divided into various training and testing sets. The training set is used to determine the best classifier for a given gene set while the testing set is used to evaluate their performances. Generally, this gives better classification performance in practice, but the computational time may be prohibited if one wishes to try every single possible combination from the gene set. Moreover, the direct interpretation of the gene ranking (GR) is difficult due to the variation of GRs across different CV subsamples. As an illustration, Table I shows the top three ranked genes which are selected by \( t \)-test from different subsamples generated by 10-fold CV. It can be seen in Table 1 that the genes ranked by the \( t \)-test method in general vary considerably across CV subsamples.

To overcome these limitations, we propose in this paper an approach based on the ridge partial least squares [21] (RPLS) with recursive feature elimination (RFE), and novel Brier and occurrence probability (OP) measures, or RPLS-RFE method in short. The RPLS is commonly used in classification problems, and it combines the ridge logistic regression (LR) and the PLS [22] to reduce the large variance in estimation. However, when it is applied to the detection of cancer-related genes, one has to test all combinations according to certain criteria, such as misclassification error derived from the RPLS. Unfortunately, the complexity could be very high because the total number of genes is generally large and the number of cancer-related genes is unknown. Hence, an efficient approach that can choose the desired genes systematically is required. A possible candidate is the RFE, which is a backward elimination procedure. In the context of microarray analysis, it can be used to efficiently remove genes with least contribution to a classifier and to reduce the computational time for model selection to an acceptable level.

The proposed method possesses three distinct and desirable features.

1) It employs the RPLS method so that the correlation between the genes from different networks can be taken into account. Consequently, it is able to explore the pathway dependencies of the genes involved. Together with the RFE, the variance of the regression coefficients can be significantly reduced, while considerably reducing the computational time in finding the relevant genes.

2) Novel Brier measure: Although combining RPLS and RFE helps to extract the correlation among the genes, it is still insufficient to detect significant genes for the identification of GRNs. In conventional RFE, misclassification error is often employed as a selection criterion for classification problem. However, such hard decision metric is not a good choice for our purpose because some useful information regarding the gene-gene correlation may be lost due to the quantization of the metric. Therefore, we proposed to employ the Brier score (BS) which is a soft decision metric. The advantage of the BS is that it is continuous, and hence it is less prone to quantization effect of conventional misclassification error and provides a more accurate estimation of the prediction model. Though RPLS has been reported before for classification, the incorporation of RFE with Brier error measure and its application to preliminary GRN identification is to our best knowledge new.

3) Novel OP measure: As mentioned earlier, although the conventional CV approach was proved to improve classification performance in many studies that focus on classification, the extracted GRs are generally not consistent say for different portion of data samples as shown in Table I. Hence, we adopt the recent two-nested CV procedure to randomly partition data samples in order to explore the variations of the genes chosen by the RPLS-RFE algorithm on different portions of data samples. Another advantage of the two-nested CV procedure is that it avoids the problem of overfitting (also known as selection bias) when using the same CV loop for both parameter tuning and performance evaluation [23]. Moreover, we propose a new measure called the OP, which counts the occurrence of a given gene on the chosen gene sets across different CV subsamples. This provides a very effective criterion for selecting correlated genes with potentially biological significance, rather than just reducing the size of the gene sets to boost the recognition rate using different gene sets. It is an important advantage that allows us to detect different GRNs from the correlated genes, which consistently appear in the ranking lists of

<table>
<thead>
<tr>
<th>Subsamples generated by 10-fold cross validation</th>
<th>Gene Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1772 780 1771</td>
</tr>
<tr>
<td>2</td>
<td>1772 1771 1582</td>
</tr>
<tr>
<td>3</td>
<td>1582 1772 249</td>
</tr>
<tr>
<td>4</td>
<td>1772 515 625</td>
</tr>
<tr>
<td>5</td>
<td>1772 513 1060</td>
</tr>
<tr>
<td>6</td>
<td>1582 1772 513</td>
</tr>
<tr>
<td>7</td>
<td>249 1582 1772</td>
</tr>
<tr>
<td>8</td>
<td>1772 1582 249</td>
</tr>
<tr>
<td>9</td>
<td>249 1772 245</td>
</tr>
<tr>
<td>10</td>
<td>1772 1582 780</td>
</tr>
</tbody>
</table>
the CV subsamples through public domain gene network database.

By combining the good performance of the proposed RPLS-RFE method and the novel Brier and OP-based measures, it will be shown later by our experimental results on three publicly available cancer data sets namely, the leukemia [9], colon [24], and prostate [25], that very stable gene sets from different but relevant GRNs can be identified using a two-nested CV procedure. Moreover, most of them are found to be of biological significance after carefully checking with related literature based on biological experiments. To further substantiate the statistical significance of the extracted gene sets, a model that assumes all the genes are solely picked by random is used to show that these gene sets appearing in the multiple CV trials is not a result of statistical variation. More precisely, 43, 40, and 38 out of 50 extracted genes using the proposed OP criterion for the leukemia, colon, and prostate data sets, respectively, are found to be significant at 5% significance level, which is similar in spirit to the commonly used p-value. This provides strong statistical significance of the gene sets extracted. Furthermore, almost all of these genes and GRNs cannot be revealed by conventional methods which focus mainly on the classification performance. These findings suggest that the proposed method is capable of detecting preliminarily genes and associated GRNs from the microarray cancer data, which may serve as a useful tool for preliminary detection of genes and their associated GRNs associated with a particular disease from microarray or similar data for further biological studies. Finally, it should be noted that the present work is different from the gene network identification from time course gene microarray data, which may allow one to study the detailed regulatory mechanism of the networks. The main objective of this work is to employ “one-time point” labeled microarray data to identify stable gene sets which may be related to a given disease with high statistical significance, so called “targets,” and to perform preliminary identification of GRNs or pathways with the aid of public domain gene network database. The ultimate goal is to conduct detailed biological experiments on these targets to reveal more clearly the underlying regulatory mechanism.

The paper is organized as follows. Previous works on gene identification are briefly reviewed in Section II. Section III is devoted to the proposed method. In Section IV, we provide a comprehensive evaluation of the proposed method with other conventional algorithms and demonstrate new results on preliminary identification of cancer-related genes and their associated GRNs. Finally, conclusion is drawn in Section V.

II. PREVIOUS WORKS

A microarray sample consists of gene expression intensities collected for a large number of genes and its associated class label or observation, which can be binary or multiclass. In this paper, we focus on the analysis of commonly encountered two-class microarray data. More specifically, consider an observation \( y = 0, 1 \), which depends on the expressions of \( J \) genes of a subject given by an \((J \times 1)\) variable vector \( x = [x_1, x_2, \ldots, x_J]^T \), where the superscript \( T \) denotes matrix transpose. Usually, due to large dynamic range of gene expressions, data preprocessing such as logarithmic transformation will be performed on \( x \).

Suppose further that we have the gene expressions and the associated labels of \( N \) subjects or samples. Then, the \( N \) observations of \( y \) and its corresponding gene expressions can be grouped into an \((N \times 1)\) vector \( Y = [y_1, y_2, \ldots, y_N]^T \) and a \((N \times J)\) matrix \( X = [x_1, x_2, \ldots, x_N]^T \), respectively, where \( x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,J}]^T, i = 1, \ldots, N \), are the gene expressions of the \( i \)th subject. Moreover, \( X \) is centered and standardized as

\[
\overline{x}_{i,j} = \frac{(x_{i,j} - \mu_j)}{\sigma_j}
\]  

where \( \mu_j = \frac{1}{N} \sum_{i=1}^{N} x_{i,j} \) and \( \sigma_j \) are, respectively, the mean and standard deviation of the \( j \)th gene expression given by

\[
\sigma_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_{i,j} - \mu_j)^2}.
\]  

For the sake of presentation, we denote \( \overline{X} \) as the gene expression matrix after centering and standardization, and its \((i,j)\) entry is given by (1).

In the filter-based approach, the ranking of a gene is determined independently by calculating the discrimination power of each gene according to certain measure and sorting them in a descending order. Four of the most commonly used filter approaches for gene selection are S2N, t-test, PC, and FC. In these methods, the gene expression matrix \( \overline{X} \) is further divided into two subclasses \( \overline{X}^- \) and \( \overline{X}^+ \) for class 0 and class 1, i.e., the samples with label \( y = 0 \) and \( y = 1 \), respectively. The S2N method measures the mean separation of the two classes for a particular gene with respect to their spread as follows:

\[
SNR(j) = \frac{|\mu^-_j - \mu^+_j|}{\sigma^-_j + \sigma^+_j}
\]  

where \( \mu^-_j \) and \( \mu^+_j \) (\( \sigma^-_j \) and \( \sigma^+_j \)) are, respectively, the mean and standard deviation of the \( j \)th gene expression of \( \overline{X}^- (\overline{X}^+) \). The t-test method further takes the number of samples in each class into account in computing the spread

\[
T(j) = \frac{|\mu^+_j - \mu^-_j|}{\sqrt{(\sigma^+_j)^2/N^+ - (\sigma^-_j)^2/N^-}}
\]  

where \( N^- \) and \( N^+ \) are the total number of samples in classes 0 and 1, respectively. The PC method uses a similar but different weighting as follows:

\[
PC(j) = \frac{\mu^+_j - \mu^-_j}{\sqrt{\frac{N^- N^+}{N^- + N^+} \sum_{i=1}^{N} (x_{i,j} - \mu_j)^2}}.
\]  

On the other hand, the FC method measures the normalized Euclidean distance between the centers of the two classes as

\[
fc(j) = \frac{(\mu^+_j - \mu^-_j)^2}{(\sigma^+_j)^2 + (\sigma^-_j)^2}.
\]  


The identification of relevant genes for a particular disease from labeled gene microarray data is then performed by ranking the genes according to the scores produced by one or more of these measures. After that, a subset of \( sJ \), \( 0 \leq s \ll 1 \), ranked gene indices, \( S_{\text{sub}} \), in descending order is obtained

\[
S_{\text{sub}} = \{1 \leq j \leq J : |d(j)| \text{ is the first } sJ \text{ largest of all}\}
\]

where \( \lceil . \rceil \) is the ceiling operator, and \( d(j) \) can be either \( \text{SNR}(j), T(j), f_c(j), PC(j), \) or some combinations of them. The parameter \( s \) is usually chosen as a value much smaller than 1 to retain a small subset of sorted genes. Generally, these selected genes in \( S_{\text{sub}} \) are considered to be more statistically significant and they are retained for classification or further studies. However, a major drawback of the filter approach is that the correlation between the genes is ignored, and hence some biological significant genes may be rejected.

In the wrapper-based approach, classification accuracy is used as the scoring metric to grade the relevance of the chosen genes. First, a search algorithm is used to search through the space of possible subset of genes. Then, the classifier is invoked to evaluate the classification accuracy of the chosen subset. Consequently, the subset of genes that gives the best classification accuracy is retained for classification.

Motivated by the simplicity of the filter-based approach and the good classification accuracy of the wrapper-based approach, a hybrid approach was proposed in [16]. A filter algorithm is first invoked on the whole set of genes to retain only a small subset of relevant genes. Then, the wrapper algorithm is invoked on the retained gene subset of a much smaller size. Hence, the computation speed of the wrapper algorithm can be further improved. In particular, a multiple-filter-multiple-wrapper algorithm was proposed in [17], which combines the use of different filter metrics so that more relevant genes are retained. Moreover, multiple wrappers are used to choose the genes according to a consensus established among several classifiers. Hence, the identified genes generally have a mixture of characteristics that gives good classification accuracy for the included classifiers.

However, the selection process in the aforementioned approaches either do not take into account correlation of genes or focusing only on a few key genes for classification purpose. Therefore, they may not able to identify relevant genes from different GRNs. Next, we will present the proposed approach which aims at identifying a more comprehensive set of genes from different GRNs, without sacrificing the classification performance. These genes can serve as potential targets for investigating more detailed regulatory actions between the GRNs using biological experiments and time course data.

### III. Proposed Approach

The proposed approach can be summarized as follows: First, a two-nested CV procedure is used to divide the data set into subsamples for training and testing. An advantage of the two-nested CV procedure is that it avoids the selection and optimization bias found in the conventional CV procedure [20], [23]. Second, the RPLS and the RFE algorithms are invoked recursively so as to eliminate irrelevant genes from the selected gene subsets using the BS. Finally, the proposed OP-based measure is used to rank the genes obtained across different CV subsamples. Table II summarizes the key implementation steps of the proposed method. For better exposition and understanding of the RPLS algorithm, it is summarized in Section III-A below.

#### A. Ridge Partial Least Squares

The RPLS method [21] combines PLS with LR and ridge regression to reduce the variance of the regression estimate. LR is a type of generalized linear model used for binary regression. It only relies on the conditional class probability \( P(y|x) \), or equivalently the conditional expectation of \( Y \) given...
\[ x, \text{ and so it does not require the usual normal assumption on the distribution. This makes it less sensitive to model misspecification [26]. On the other hand, PLS is an efficient dimension reduction technique which can be applied to LR to reduce the large variance resulting from the high dimension and small number of samples problem. In RPLS, ridge regression is further employed with PLS to reduce the variance of the estimates.} \]

More precisely, for binary regression, LR models the conditional class probability \( P(y = 1|x) \) for \( x \) in class 1 as the following function of the explanatory variables \( x \):

\[
P(y = 1|x) = \frac{e^{\eta(\alpha, \beta, x)}}{1 + e^{\eta(\alpha, \beta, x)}} = h(\eta(\alpha, \beta, x)) = p(\alpha, \beta, x)
\]

where the link function of the GLM, \( h(\eta) \), is chosen as the logistic function \( e^\eta/(1 + e^\eta) \) and \( \eta(\alpha, \beta, x) \) is chosen as a linear predictor in terms of \( x \) with regression coefficients \( \beta = [\beta_1, \beta_2, \ldots, \beta_J]^T \) and intercept \( \alpha \):

\[
\eta(\alpha, \beta, x) = \alpha + x^T \beta.
\]

Using (8) and (9), one can verify that \( \ln(P(y = 1|x)) - \ln(1 - P(y = 1|x)) = \alpha + \beta^T x \). Since it is a binary regression, \( 1 - P(y = 1|x) \) represents the conditional class probability for \( x \) to be in class 0. The unknown parameters \( \alpha \) and \( \beta \) are usually estimated by maximum likelihood method, which can be computed numerically by using the iteratively reweighted least squares (IRWLS) method. More precisely, given the observations \( Y^{TRAIN} \) and processed input gene expression matrix \( X^{TRAIN} \) of a training set \( S^{TRAIN} \), the likelihood function can be written as

\[
\log(L(\alpha, \beta|S^{TRAIN})) = \sum_{i = 1}^{N^{\text{TRAIN}}} \log(P(\alpha, \beta, x_i)) \]

where \( N^{\text{TRAIN}} \) is the number of training samples, and we have assumed that the samples are independent. The regression coefficients can be estimated by maximizing the likelihood function or equivalently the log-likelihood function as follows:

\[
\ln(L) = \sum_{i = 1}^{N^{\text{TRAIN}}} \log(p_1(\alpha, \beta) + (1 - y_i) \log(1 - p_1(\alpha, \beta))]
\]

where for notation convenience, we have used \( p_1(\alpha, \beta) \) for \( p(\alpha, \beta, x_i) \), \( \ln(L) \) for \( \log(L(\alpha, \beta|S^{TRAIN})) \), \( N^{\text{TRAIN}} \) for \( N^{\text{TRAIN}} \) and \( Y \) for \( Y^{TRAIN} \). To simplify notation, we further define

\[
\hat{X} = [1_{N'}, \ X^{TRAIN}] \quad \text{and} \quad \gamma = [\alpha, \beta^T]^T.
\]

where \( 1_{N'} \) denotes an \((N' \times 1)\) vector with all entries equal to ones. As shown in Appendix A, the desired solution of maximizing the log-likelihood in (10) satisfies

\[
(X^T \hat{W} \hat{X}) \gamma \approx (X^T \gamma + \hat{W} \hat{X} \gamma)
\]

where \( \hat{W} = \text{diag}\{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_{N'}\} \), \( \hat{w}_i = p_i(1 - \hat{p}_i) \), and \( \hat{p}_i = \hat{p}_i(\gamma) = p_1(\alpha, \beta) \). By solving (12) and using the solution as the new estimate \( \gamma = [\hat{\alpha}, \hat{\beta}^T]^T \) repeatedly, one obtains the IRWLS algorithm for solving the regression coefficients. Usually, the iteration stops when a maximum number of iterations is reached or when the change in successive \( \gamma \)'s is sufficiently small. From (12), one can also notice that

\[
(X^T \hat{W} \hat{X}) \gamma \approx (X^T \gamma + \hat{W} \hat{X} \gamma) = (X^T \gamma) Z
\]

where \( Z = X \gamma + \hat{W}^{-1}(Y - \hat{p}) \) is the logistic pseudo variable and is recognized as a better approximation to \( X \gamma \) and hence \( h(\gamma) \).

In underdetermined or nearly singular problems, regularization should be applied to tackle the singularity problem of the matrix \( X^T \hat{W} \hat{X} \) in (12) by adding to it a small matrix \( \lambda R \). This gives

\[
(X^T \hat{W} \hat{X} + \lambda R) \gamma \approx X^T (Y - \hat{p} + \hat{W} \hat{X} \gamma)
\]

where \( \lambda \) is the positive regularization parameter and \( R \) is a diagonal matrix with \( R_{11} = 0 \) and \( R_{jj} = 1, j = 1, \ldots, J \). Equation (14) can be repeatedly used to refine the approximation \( \gamma \) obtained in the previous iteration, which constitutes the desired IRWLS algorithm with regularization. In [21], the regularization parameter is chosen by the Bayesian information criterion. Whereas in this work, CV is used to fine-tune this parameter from the training data.

Though regularization reduces the estimation variance and hence gives a good estimator when \( \hat{X}^T \hat{W} \hat{X} \) is ill-conditioned, it is also biased. To improve upon the regression vector \( \gamma = [\alpha, \beta^T]^T \), it is suggested in [21] to apply PLS fit of \( \hat{X} \gamma \) to \( \hat{Z} \) obtained in the previous step

\[
\hat{Z} = X \gamma + \hat{W}^{-1}(Y - \hat{p}(\gamma))
\]

where \( \hat{\gamma} = [\hat{\alpha}, \hat{\beta}^T]^T \), \( \hat{W}^{-1} \) and \( \hat{p}(\gamma) = [p_1(\gamma), \ldots, p_{N'}(\gamma)]^T \) are computed using the results of the penalized LR above. This is because, as suggested in (13), \( Z \approx X \gamma \), where \( \gamma \) is the true regression vector. An important advantage of using the pseudo variable \( \hat{Z} \) is that it combines the PLS dimension-reduction step with the previous ridge regularization step. The former removes noise in less important dimensions by projecting \( \hat{Z} \) onto a low dimensional subspace spanned by the principal components (PCs) and hence de-noises \( \hat{Z} \) through thresholding the unimportant components. More precisely, \( \hat{Z} \) is written as

\[
\hat{Z} = \hat{X} \gamma + E'
\]

where \( E' \) is the error which is assumed to be normal distributed. In the PLS, \( \hat{X} \) and \( \hat{Z} \) are related to some latent variables \( \Gamma = [\bar{t}^{(1)}, \ldots, \bar{t}^{(M)}] \) and \( \hat{\Gamma} = [\hat{t}^{(1)}, \ldots, \hat{t}^{(M)}] \) of the PC directions. We are not concerned about the covariance of \( \hat{X} \) and the covariance of \( \hat{Z} \) themselves, but the cross-variance between \( \hat{X} \) and \( \hat{Z} \), i.e., \( \hat{X} \hat{Z} \). More precisely, we use two different bases \( \hat{\Gamma} \) and \( \hat{\Gamma} \) to, respectively, decompose \( \hat{X} \) and \( \hat{Z} \) as follows:

\[
\hat{X} = \sum_{m=1}^{M} t^{(m)} p^{(m)} = E \Gamma P^T + E
\]

\[
\hat{Z} = \sum_{m=1}^{M} \hat{t}^{(m)} c^{(m)} = \hat{\Gamma} \hat{C}^T + E
\]

where \( \Gamma = [\bar{t}^{(1)}, \ldots, \bar{t}^{(M)}]^T \), \( P = [p^{(1)}, \ldots, p^{(M)}]^T \), \( \bar{t}^{(m)} \) is an \((N \times 1)\) vector, \( p^{(m)} \) is an \((J \times 1)\) vector, \( E \) is the residual.
of $\hat{X}$, and $M$ is the number of chosen PCs. Similarly, $F$ is the residual of $Z$ and $C = [c^{(1)}, \ldots, c^{(M)}]$. $\Gamma$ and $\tilde{\Gamma}$ are also called the score matrices containing the directions of the PCs, while the matrices $P$ and $C$ are called the loadings of the PCs.

In [21], $M$ is chosen by minimizing the misclassification error in the CV. Note, if $t^{(m)} = p^{(m)}$, we have the conventional PC analysis of $\hat{X}$. However, such an expansion will not be able to explain $\tilde{Z}$ effectively at the same time and may lead to inferior performance.

In PLS, $t^{(m)}$ and $\tilde{t}^{(m)}$ are determined successively from the data. At the $m$th iteration, one can work only with the regression vector $\beta^{(m)}$ can be ignored the intercept term $\omega$. The predicted conditional class probability for the testing data $\hat{c}(\tilde{X}^{(test)})$ is obtained as

$$
\hat{c}(\tilde{X}^{(test)}) = \frac{e^{\hat{X}^{(test)}\beta^{(m)}}}{1 + e^{\hat{X}^{(test)}\beta^{(m)}}}
$$

Using the Lagrange multiplier method, $\omega^{(m)}$ and $q^{(m)}$ can be solved as an eigenvalue problem, as described in Appendix B. The final regression vector $\gamma_{PLS}$, which is defined similarly as in (11), is obtained as

$$
\gamma_{PLS} = \tilde{\Phi}(P^T\tilde{\Phi})^+C^T
$$

where $\tilde{\Phi} = [\omega^{(1)}, \ldots, \omega^{(M)}]$, $\tilde{P} = [p^{(1)}, \ldots, p^{(M)}]^T$, $C = [c^{(1)}, \ldots, c^{(M)}]$ and $(\cdot)^+$ denotes the pseudo inverse operator. For notation convenience, we drop the subscript and consider $\gamma$ as the final regression vector in subsequent sections.

If the data is centered, then the intercept is equal to zero, and one can work only with the regression vector $\beta$. Alternatively, one can ignore the intercept term $\alpha$ and work with $\beta$ only, $\alpha$ can then be computed as

$$
\alpha = \tilde{X}^T\tilde{W}1_{N_{test}} - (Z^T\tilde{W}1_{N_{test}})^T\beta.
$$

The predicted conditional class probability for the testing data matrix $X_{test}$ with $N_{test}$ samples is given by

$$
\hat{p} = h(\tilde{\eta})
$$

where $\tilde{\eta} = \alpha 1_{N_{test}} + \tilde{X}_{test}\beta$. Here, a two-nested CV [20] may be preferred over the single-loop CV used in [21] because using the same CV loop for both parameter tuning and performance evaluation may introduce overfitting. Moreover, if a subset of genes is required, then direct truncation from the whole gene sets according to some benchmark criteria may be undesirable. A better approach is to employ the wrapper-based approach where the classifier should be evaluated for the given gene set only. However, testing all such combinations will be computationally prohibitive, and hence RFE has to be employed.

### B. Recursive Feature Elimination With Brier Error Measure

The use of RFE with support vector machine for efficient feature selection was first proposed in [15]. Motivated by the effectiveness of RFE, we now extend the RPLS algorithm to include RFE. In RFE, feature selection is performed by eliminating variables that contribute least to the classifier recursively according to certain performance criterion until a minimum gene number is obtained as in many greedy classifiers or a minimum acceptable classifier performance is achieved.

More specifically, in [15], each variable in the regression vector $\beta$ obtained in (21) corresponds to the $q$th gene, and the regression coefficients $\beta_j$ are ranked in descending order of their squared magnitudes. The intercept obtained in (22) is ignored because it does not correspond to any of the genes. Then, the irrelevant genes are eliminated at each iterations. Let $q = 1, 2, \ldots, Q$ denotes the number or index of iterations with the initial gene index set at the $q$th iteration given by $S^{(q-1)}$ having $J^{(q-1)}$ elements. A classifier is then designed using the gene set $S^{(q-1)}$. By ranking the resulting regression coefficients according to their magnitudes and retaining only $s^{(q)}$ fraction of the original gene sets in $S^{(q-1)}$, we obtain the new reduced gene set $S^{(q)}$ at the $q$th iteration as follows:

$$
S^{(q)} = \left\{ j \in S^{(q-1)} : \beta_j \text{ is the first } \left[ s^{(q)} J^{(q-1)} \right] \right\},
$$

where $s^{(q)}$ is usually initialized to include all the genes in the first iteration, i.e., $S^{(1)} = \{1 \leq j \leq J\}$, where $J$ is the total number of genes of the given samples. The parameter $s^{(q)}$ is called the step size for variable selection at the $q$th iteration. In [15], a fixed step size $s^{(q)} = 1/2$ is chosen, and half of the genes will be eliminated. The number of genes retained is

$$
J^{(q)} = s^{(q)} J^{(q-1)}.
$$

Hence, the total number of genes retained when the RFE algorithm terminates is $J^{(Q)} = s^{(1)} J^{(1)} \ldots s^{(Q-1)} J^{(Q-1)} J$. Since $0 < s^{(q)} < 1$, the number of retained genes are usually much smaller than the original number of genes, i.e., $J^{(Q)} \ll J$. The classification accuracy at each iteration is used as the performance criterion for various gene subsets generated. The RFE continues to eliminate half of the genes and terminates when only one gene is left or no genes can be further eliminated.

Unlike [15], we adopt a variable step size which adapts to the performance of the classifiers in our proposed approach. More precisely, at each iteration, the RPLS algorithm is invoked on $Y$ and $X_{sub}$, where $X_{sub}$ is the standardized gene expression matrix that contains only the selected gene indices given by $S^{(q)}$. An inner CV is employed to determine the required regularization for the given gene set. Then, the regression vector
Fig. 1. Decision network diagram for the recursive feature elimination (RFE).

\[
\beta \text{ is obtained and sorted as in (24). Afterwards, the estimated performance measure for the } q\text{th iteration, denoted by } BS^{(q)}, \text{ is obtained from the RPLS-based classifier. Here, the performance measure we employ is the BS [27], which will be discussed later. If the current estimated } BS, BS^{(q)}, \text{ is smaller, the step size for the next iteration is halved}
\]

\[
s^{(q+1)} = \left(\frac{1}{2}\right)s^{(q)} \text{ if } BS^{(q)} < BS^{(q-1)}
\]

\[
(26)
\]

so that the number of genes eliminated is doubled to speed up the elimination process. Otherwise, the same iteration is repeated, except that the step size is increased to \(s^{(q)} = 1 - \frac{1}{2}(1 - s^{(q)})/2\) so that the fraction of genes to be eliminated is reduced to increase the sampling in search of a better gene set. Fig. 1 shows the decision network diagram of the proposed RFE approach.

The proposed RPLS-RFE algorithm continues until only one gene is left or no gene can be further eliminated. The concept of BS, which is a soft decision metric, is adopted in this study. Unlike the hard decision as in the conventional misclassification error, it is less prone to thresholding effect and provides more precise information to evaluate the reliability of the prediction model. In digital communications applications, it has been shown that using soft decision metrics usually leads to better performance than hard decision metrics. A classical example is the Viterbi decoder [28] with soft decision output. In our proposed RPLS-RFE algorithm, the BS is expressed in terms of the conditional class probability estimate as follows:

\[
BS = \sum_{i=1}^{N_{\text{TEST}}} \left( y(i) - h(\alpha + \mathbf{X}_i^T \beta) \right)^2
\]

\[
(27)
\]

where \(y(i)\) is the \(i\)th observation of the testing set \(Y_{\text{TEST} - \text{INT}} = [y(1), y(2), \ldots, y(N_{\text{TEST}})]^T\), \(x(i)\) is the \(i\)th sample of the testing samples \(X_{\text{TEST} - \text{INT}} = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N_{\text{TEST}})]^T\), \(N_{\text{TEST}}\) is the total number of samples in the testing set, \(\alpha\) is the intercept in (22), \(\beta\) is the regression vector in (21), and \(h(\eta)\) is the logistic function in (8). For notation convenience, we have dropped the dependency of BS on the regression vector \(\beta\).

In general model selection problem using stepwise linear regression techniques such as the RFE, the best model can be obtained by selecting the set of variables with some appropriate error measures such as minimum mean squared error [29]. Similar idea applies to the proposed approach, where adding irrelevant genes to or removing relevant genes from the retained gene subset would increase the BS in (27). Therefore, we expect that the resultant BS attains its minimum value when the best model is selected. Next, the details of the two-nested CV for parameter and model selection will be described.

C. Two-Nested Cross Validation

CV is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is widely used to estimate the practical performance of a model for prediction, classification, and related applications. It is usually divided into several rounds, and each of them partitions the sample data into complementary subsets, performing the analysis on one subset, called the training set, and evaluating the performance of the analysis on the other, called the testing set. To reduce statistical variability, multiple rounds of CV are performed using different partitions, and the validation results or performance are averaged over all rounds. CV can also be used in parameter tuning by selecting the best parameters from a given set. This is widely used in determining the regularization parameters in ridge or other regression techniques. In repeated random subsampling, the partitions are randomly chosen.

In \(K - \text{fold} CV\), the original sample is randomly partitioned into \(K\) subsamples. Of the \(K\) subsamples, a single subsample is retained as the validation data for testing the model, and the remaining \(K - 1\) subsamples are used as training data. The CV process is then repeated \(K\) times (the folds), with each of the \(K\) subsamples used exactly once as the testing data. The \(K\) results from the folds are then averaged, or otherwise combined, to produce the estimated performance. Its advantage over repeated random subsampling is that all observations are used for both training and testing, and each observation is used for validation exactly once. When \(K\) is equal to \(N\), the total number of samples, it is referred to as the leave-one-out CV (LOOCV). However, it is usually very expensive from a computational point of view because of the large number of training processes involved. Both the \(10 - \text{fold} CV\) and the LOOCV are widely used in microarray studies.

The two-nested CV procedure adopted in this work is a recent tool that has received considerable attention in the bioinformatics community [20], [23], [30], [31]. In classification, an important advantage of the nested CV procedure is that it overcomes the limitation of the optimistic and selection biases [20], [23] often found in using conventional CV.
procedure in joint parameter tuning and performance evaluation. In two-nested CV, an internal CV is used for parameter tuning, while the external CV is used for evaluating the classification performance.

In the proposed RPLS-RFE method with two-nested CV, the external CV is used to generate different subsamples so that different gene subsets can be found and the variability/predicting power can be studied, while the inner CV will be used to determine the regularization parameter $\lambda$ in (14) and the number of PCs $M$ in (17). Moreover, we apply stratification on the CV so that each subsample has an equal proportion of subjects belonging to different classes, e.g., 60% cancer and 40% normal samples. This procedure helps to reduce bias due to unbalanced samples. This is different from the RPLS method in [21], which uses a single-loop CV procedure for parameter tuning and evaluating the classification performance. While in [15], the single-loop CV procedure is used for feature selection, parameter tuning, and performance evaluation of classification.

More specifically, consider a $K$-fold CV procedure, where the observation $Y$ and gene expression matrix $X$ is partitioned into $K$ subsamples. At the $k$th loop, $k = 1, 2, \ldots, K$, the testing set is formed by the $k$th subsample $(Y_k, X_k)$ such that

$$Y_{\text{TEST}} = Y_k = [y_{ka+1}, y_{ka+2}, \ldots, y_{(k+1)a}]^T$$

$$X_{\text{TEST}} = X_k = [x_{ka+1}, x_{ka+2}, \ldots, x_{(k+1)a}]^T$$

where $a = \lceil N/K \rceil$, $\lceil N/K \rceil$ is the integer part of $N/K$, and $N$ is the total number of samples. If $N$ is not divisible by $K$ and the reminder is $A$, the remaining samples will be distributed among the first $A$ subsamples so that they have one extra sample more than the remaining $K-A$ subsamples. The training set $(Y_{\text{TRAIN}}, X_{\text{TRAIN}})$ is formed by the remaining $K-1$ subsamples as

$$Y_{\text{TRAIN}} = [Y_{1}^T, Y_{2}^T, \ldots, Y_{k-1}^T, Y_{k+1}^T, \ldots, Y_{K}^T]^T$$

$$X_{\text{TRAIN}} = [X_{1}^T, X_{2}^T, \ldots, X_{k-1}^T, X_{k+1}^T, \ldots, X_{K}^T]^T.$$ (31)

Centering in (1) and standardization in (2) are invoked on $X_{\text{TRAIN}}$ and $X_{\text{TEST}}$ using the mean value and standard deviation of $X_{\text{TRAIN}}$ to obtain $\overline{X_{\text{TRAIN}}}$ and $\overline{X_{\text{TEST}}}$, respectively. The algorithm to be evaluated is then invoked using the training set for learning and the testing set for evaluating its performance.

For the two-nested CV adopted in our proposed approach, the samples are first partitioned into $K$ subsamples. At the external CV loop, the training set $(Y_{\text{TRAIN\_EXT}}, X_{\text{TRAIN\_EXT}})$ and testing set $(Y_{\text{TEST\_EXT}}, X_{\text{TEST\_EXT}})$ are the same as those in (28)–(31). The external $K$-fold CV loop is used to evaluate the performance of the RPLS-RFE algorithm and to obtain different selected gene subsets from the subsamples. In the internal $K-1$ fold CV loop, the training set is given by

$$Y_{\text{TRAIN\_INT}} = [Y_{1}^T, Y_{2}^T, \ldots, Y_{k-2}^T, Y_{k+1}^T, \ldots, Y_{K}^T]^T$$

$$X_{\text{TRAIN\_INT}} = [X_{1}^T, X_{2}^T, \ldots, X_{k-2}^T, X_{k+1}^T, \ldots, X_{K}^T]^T.$$ (33)

so that both the $(k-1)$th and the $k$th subsamples are omitted from the training set. The testing set is given by

$$Y_{\text{TEST\_INT}} = Y_{k-1}, \quad X_{\text{TEST\_INT}} = X_{k-1}. \quad (34)$$

Centering and standardization are then performed similar to those for the external training and testing sets. The internal CV loops are then used for parameter tuning of the RPLS-RFE algorithm. The regularization parameter $\lambda$ in (14) and the number of PCs $M$ in (17) are chosen to minimize the estimated error returned by the RPLS-RFE algorithm in the internal CV loops. $\lambda$ and $M$ are confined to a set of possible values and a 2-D grid search is used to choose the best parameters that yield the minimum error measure. This offers some protection against local minima [32]. Moreover, to speed up the process, we adopted an iterative grid search approach that is similar to [33].

Now, we discuss the error measure used for parameter tuning. Conventionally, the classification error and the area under the receiver operation characteristic curve are used as error measures for parameter tuning in many classification algorithms. These measures are based on hard decision metrics, i.e., the metric is discrete, such as true positive (TP), false positive (FP), and so on. On the other hand, in the proposed approach, we adopt the concept of BS [27], which is a soft decision metric. As mentioned earlier, it provides more precise information to ensure the reliability of the prediction model.

### D. Occurrence Probability Measure for Gene Ranking

As mentioned earlier, the high dimensionality, large biological variability, and small number of samples in gene microarray analysis create extremely large statistical variations in conventional regression techniques and hence complicated the detailed analysis of gene microarray data and related biological data. Therefore, conventional approaches employ gene selection and work with very few gene sets to boost the performance of the classifiers. However, this approach sacrifices the ability to identify related genes in the GRNs as studied in this paper, through which the GRNs can be inferred from public domain gene network database. Another major difficulty in identifying these genes is that, after CV, many methods generate gene sets with very large variability, and hence it is difficult to give a reliable ranking of the genes.

In our approach, two-nested CV using RPLS-RFE with Brier measure helps us to reduce significantly the statistical and biological variations. If the gene sets are stable, then the relevant genes should in principle appear in different external CV rounds. Hence, it is natural to employ the $\text{OP}$ of a given gene in the final gene set of different CV rounds as a measure of its importance.

More specifically, consider a chosen subset of gene indices $S^{(Q)}$ estimated from a subsample in CV defined in (39) to (42). We are given $K$ subsamples, and the set of gene indices chosen for all subsamples is $S = \{S^{(Q)}_1, S^{(Q)}_2, \ldots, S^{(Q)}_K\}$, where $S^{(Q)}_k$ is the chosen subset of gene indices obtained from the $k$th subsample. In the subsamples, only $J^{(Q)}$ genes are retained by the RPLS-RFE algorithm. Also, $J^{(Q)} \ll J$ and hence most of the irrelevant genes are omitted from subsamples. The proposed
**TABLE III**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Samples</th>
<th>Number of Genes/Probes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Type 0</td>
<td>Type 1</td>
</tr>
<tr>
<td>Leukemia (A)</td>
<td>38</td>
<td>27</td>
<td>11</td>
</tr>
<tr>
<td>Colon</td>
<td>62</td>
<td>40</td>
<td>22</td>
</tr>
<tr>
<td>Prostate</td>
<td>102</td>
<td>50</td>
<td>52</td>
</tr>
</tbody>
</table>

Identification of cancer-related genes and GRNs

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Identification of cancer-related genes and GRNs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia (B)</td>
<td>72</td>
</tr>
<tr>
<td>Colon</td>
<td>62</td>
</tr>
<tr>
<td>Prostate</td>
<td>102</td>
</tr>
</tbody>
</table>

**OP** measure for ranking the gene indices across different subsamples is given by

\[ OP(j) = \frac{k_j}{K} \]  

where \( k_j \) is the number of occurrence of the \( j \)th gene among \( K \) subsamples. Since only \( J(Q) \) genes are retained in the subsamples, the \( OP \) of the irrelevant genes eliminated by RFE will vanish since \( k_j = 0 \). As this measure extracts a group of genes that bear high consistency across “different subsamples,” they are more likely to contribute to the biological disease or event under study. Moreover, such consistency will also help us to discover the interaction between important genes that may originate from different regulatory structures. By tracing the top ranked genes back to their gene pathways in public domain gene network database, there is a high hope of finding out the gene pathways or GRNs involved in that disease. From the experimental results to be presented below, we show that a number of relevant genes and GRNs can be preliminarily identified in such a manner and most of them are consistent with the findings in biological experiments.

**IV. RESULTS**

In this section, we consider the preliminary identification of cancer-related genes and their associated GRNs. Moreover, since classification performance is a conventional criterion for assessing the performance of an algorithm, we also compare the classification performance of our proposed RPLS-RFE algorithm with other conventional approaches. We consider three real cancer data sets, namely leukemia [9], colon [24], prostate [25], and they are obtained from the Kent Ridge Biomedical Data Set Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd/). A brief description of the data sets is summarized in Table III.

**A. Classification Results**

As an illustration, we compare the classification performance of our proposed approach (RPLS-RFE) and other conventional approaches reported in [30]. In the evaluation of classification performance, we adopted the two-nested CV procedure, which is a recent tool receiving considerable attention in the bioinformatics society because it overcomes the limitation of selection and optimistic bias [20], [23] found in the conventional CV procedure, say in [15] and [21]. Since all algorithms should be evaluated using the same CV procedure, we do not attempt to compare the results with [15] and [21] for fair comparison.

There are two commonly used CV procedures, namely the LOOCV and the \( 10 \)-fold CV. For fair comparison, we adopted a two-nested \( 10 \)-fold CV procedure described in [30]. The mean classification accuracy is obtained from 100 repeated runs of the two-nested \( 10 \)-fold procedure.\(^1\) The performance of all algorithms is assessed by the classification accuracy

\[ Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \]

where \( TP \) is the number of TPs, \( FN \) is the number of false negatives, \( TN \) is the number of true negatives, and \( FP \) is the number of FPs. Table IV shows the classification results for the three data sets. The results from other approaches are directly quoted from [30].

In order to study the effect of RFE and BS on RPLS, three different configurations of the proposed approach are considered using two-nested CV:

A. RPLS+RFE+BS+two-nested-CV: RFE is enabled, and only the subset of genes chosen by RFE is used for classification.

B. RPLS+BS+two-nested-CV: RFE is disabled, and all variables are retained for classification.

C. RPLS+two-nested-CV: RFE is disabled and all variables are retained for classification. Unlike the other two

\(^1\)For a \( 10 \)-fold CV, there are \( C_N^{10} \) ways of generating the subsamples, where \( C_N^k \) denotes the binomial coefficient. However, since \( C_N^{10} \) could be very large, it is a common practice as in [30], [34]–[36] to examine only part of the combinations. More precisely, combinations of different samples are chosen randomly to form partitions (subsamples), and for each combination, the two-nested \( 10 \)-fold CV procedure is invoked independently. In this section, 100 different combinations are considered for fair comparison with different classifiers reported in [30].
From Table IV, we can see that the incorporation of RFE in configuration A slightly degrades the overall classification accuracy. This is reasonable since configurations B and C use all the genes in the final classifiers. However, this may not be desirable in practice since the entire gene expressions have to be recorded, which greatly increases the practical implementation cost. If the gene set is simply truncated, then its results are usually inferior to the one using RFE, as will be shown later.

Comparing configurations B and C, we can see that the incorporation of BS in B gives lower variance in classification accuracy for the three data sets. This suggests that the continuous nature of the BS provides more precise information and is less prone to the quantizing effect of the misclassification error. Hence, it leads to higher reliability in model selection and reduces the variance of the prediction model. Also, we can see that our proposed RPLS-RFE approach generally offers highly comparable classification performance as compared with other conventional approaches.

To assess the complexity of the RPLS with and without Brier error measures, Fig. 2 shows the average computation times for the two algorithms versus the number of genes (variables). In RFE, the genes are progressively eliminated as in (26) to arrive at the desired gene sets having the best classification accuracy among all the gene sets tested. Generally, it takes an average of 4.0, 4.5, and 4.7 RFE iterations in colon, leukemia, and prostate data set to reduce the number of genes to around 100. Since the RFE can eliminate a large proportion of the irrelevant genes within a small number of iterations, the arithmetic complexity is significantly smaller than that of an exhaustive search of all possible subsets, i.e., testing the whole curve in Fig. 1. Meanwhile, the computation time is only around 1 s when the number of genes is smaller than 100. Although it takes another 16, 19, and 20 RFE iterations for colon, leukemia, and prostate, respectively, to terminate, the required computation time is only around 16–20 s. Most of the computation time originates from the first five RFE iterations which require much higher complexity.

Nevertheless, we remark that the main focus in this paper is on the preliminary identification of cancer-related genes and associated GRNs, which has not been reported using such kind of technique. We now present these novel findings.

### B. Gene and Regulatory Network Identification

Unlike the two-nested 10-fold CV used in previous section, we adopted a two-nested LOOCV procedure in this section. Since there is one (i.e., $C_N = 1$) combination of forming the CV partitions for the LOOCV, only one complete run is needed. An advantage is that it is able to obtain larger training sets among all CV procedures. This property is particularly useful for small samples because larger training sets provide more information for training the algorithm. The training set for the LOOCV has $N - 1$ samples, while the training set for tenfold CV has $0.9 N$ samples. For $N > 10$, $N - 1 > 0.9 N$. Therefore, LOOCV provides a larger training set than the tenfold CV when $N > 10$.

Both the GR and $\text{OP}$ of our proposed approach are computed according to the chosen subset of genes using the RFE as described in Section III-B. Unlike our proposed approach, the GR of the conventional approaches considered is found by invoking the algorithms on the whole data set, which may involve overfitting problem. Meanwhile, the $\text{OP}$ is found by invoking the algorithms on different CV subsamples generated by two-nested LOOCV. However, since the number of genes retained cannot be computed by the algorithms, they are usually chosen by user experience [17], [37]. As an illustration, the number of genes retained is chosen as 50 for these algorithms.

For leukemia, we checked the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [38] and identified three GRNs from the top 50 ranked genes obtained by our proposed RPLS-RFE algorithm as summarized in Table V. In particular, the functionalities of the two more important pathways are summarized below.

1) **Ribosome pathway** [39]: The Ribosome pathway is responsible for the synthesis of ribosomes, which are made to construct different kinds of proteins of the
ribosomal subunits [40] in human cells, and contributes in ribosome biogenesis. Mutation of ribosome biogenesis-related genes will cause increasing susceptibility to cancer such as leukemia [41].

2) Pathways in Cancers [42]: Different cancer may take a different pathway to become malignant, and the difference is highly variable. However, the biological endpoints ultimately reached should be shared in common. Pathways in Cancer belong to the common pathway that different type of cancers share. The pathway brings upon alternations (known as hallmark capabilities) in cell physiology that collectively dictates malignant growth.

We have also identified some cancer-related genes among the top 50 ranking genes of our proposed RPLS-RFE algorithm. Table VI shows the comparison of GR and OP between our proposed RPLS-RFE and other conventional algorithms on the cancer-related genes.

1) Metallopanstimulin 1 (HG3214-HT3391_at) is found to be a marker gene of various cancer [43]. The MPS-1 gene encodes Metallopanstimulin which is generally involved in growth factor induced response, carcinogenic ribotoxic, and the aging process [44]. Growth factors and growth factor receptors are responsible for the activation of signal transduction and the regulation of transcription. Damage to this type of genes may result in oncogenesis [44].

2) Wilm’S Tumor-Related Protein (HG3549-HT3751_at) [45] is a tumor suppressor gene. The expression of the gene suppresses the development of leukemia tumor cells [46].

3) Laminin receptor [47] (M14199_s_at) encodes the Laminin receptor, and is a biomarker of malignant transformation. Laminins are shown to be implicated in many biological processes such as cell adhesion, differentiation, migration, signaling, neurite outgrowth, and metastasis.

4) Glyceraldehyde-3-phosphate dehydrogenase (X01677_f_at) [48], is found to be involved in the regulation of apoptosis, known as programmed cell death, of leukemia cells.

5) Fibronectin 1 (HG3044-HT3742_s_at) [49]: Fibronectin is a matrix glycoprotein where the extracellular matrix influences various cellular functions such as adhesion, migration, survival, and differentiation. It was suggested to be involved in stimulating tumor cell proliferation.

6) PTMA Prothymosin alpha (M26708_s_at) [50] is found to be involved in mediating immune functions by conferring resistance to certain opportunistic infections. It is also found to be correlated to the proliferation of leukemia cells.

Overall, we can see that the pathways and genes identified are closely related to cancer and leukemia, as reported in the biology literature cited. Since the procedure of GR has been designed to reduce as much as possible the bias and variance using two-nested CV, it leads us to believe that the findings are intimately related to the underlying disease, but not
from statistical variations. Also, we note that considerable top ranked genes are associated with the ribosome and metabolic pathways. About half of the top 50 genes in our algorithm are associated with the pathways discovered. This suggests that leukemia is intimately associated with these networks. The other genes in the top 50 ranked gene list may be associated with these networks or leukemia indirectly, while possesses very high discrimination power at the same time. A possible explanation of this observation is that the strength of gene expressions for different genes may vary considerably, and therefore those intimately associated with a particular disease may have a lower gene expression than other related genes not in the main pathways [13]. Nevertheless, by exploring the consistent correlation between these genes through RPLS and OP, the proposed method appears to be able to detect those genes associated with the key pathways than those focusing entirely on a small gene list for classification purpose. In terms of classification, all the top rank genes are useful as they depicted a very high classification rate as shown in Table IV. Also, the biological variations, limited number of samples, and high dimensionality also introduce variance in their rankings.

To further substantiate the statistical significance of the extracted gene sets, we test the extracted genes against a model assuming that they are solely picked by random. Results show that 43 extracted genes using the proposed OP criterion are found to be significant at 5% significance level, which is similar in spirit to the commonly used p-value. Similar findings are obtained for the colon [24] and prostate [25] data sets, interested readers are referred to the supplementary material in [51]. As all these data sets give similar conclusions, it leads us to believe that the proposed method is capable of extracting preliminary meaningful GRNs and gene information from the gene microarray data of the respective cancer diseases studied. Of course, more studies are needed to quantify how the sample size affects the variance of the method. Studies with larger sample size are thus highly desirable in the future.
TABLE VI
RANKING AND OCCURRENCE PROBABILITY OF CANCER-RELATED GENES OF DIFFERENT APPROACHES FOR LEUKEMIA

<table>
<thead>
<tr>
<th>Genes Descriptions</th>
<th>MPS1</th>
<th>WTRP</th>
<th>LMR</th>
<th>GAPD</th>
<th>FN1</th>
<th>PTMA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ranking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPLS-RFE $^1$</td>
<td>23</td>
<td>2</td>
<td>28</td>
<td>3</td>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td>RPLS $^1$</td>
<td>125</td>
<td>12</td>
<td>9</td>
<td>6</td>
<td>244</td>
<td>1</td>
</tr>
<tr>
<td>S2N $^2$</td>
<td>2096</td>
<td>2282</td>
<td>2944</td>
<td>4770</td>
<td>2671</td>
<td>73</td>
</tr>
<tr>
<td>t-test $^3$</td>
<td>1921</td>
<td>2306</td>
<td>3127</td>
<td>4735</td>
<td>1344</td>
<td>82</td>
</tr>
<tr>
<td>FC $^2$</td>
<td>1899</td>
<td>2204</td>
<td>3032</td>
<td>4731</td>
<td>2235</td>
<td>112</td>
</tr>
<tr>
<td>PC $^2$</td>
<td>1047</td>
<td>3553</td>
<td>4479</td>
<td>4222</td>
<td>2626</td>
<td>70</td>
</tr>
</tbody>
</table>

**Occurrence Probability**

<table>
<thead>
<tr>
<th>Genes Descriptions</th>
<th>MPS1</th>
<th>WTRP</th>
<th>LMR</th>
<th>GAPD</th>
<th>FN1</th>
<th>PTMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPLS-RFE $^1$</td>
<td>0.167</td>
<td>1.000</td>
<td>0.069</td>
<td>1.000</td>
<td>0.014</td>
<td>1.000</td>
</tr>
<tr>
<td>RPLS $^1$</td>
<td>0.014</td>
<td>0.986</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S2N $^3$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>t-test $^3$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>FC $^3$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>PC $^3$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

1 Both the GR and OP of the proposed RPLS-RFE approach are calculated using the gene subsets chosen by the algorithm in the 2-nested CV procedure. The average number of retained genes across different subsamples is 33.7.
2 The gene rankings of other algorithms are obtained by invoking the algorithms on the whole dataset. Note, the RPLS in the above table is not intended for gene selection [20]. However, it is included for the sake of comparison, and genes are ranked in descending squared magnitude of the regression coefficients $\beta_j$ in (21).
3 The OPs of other algorithms are computed using the 2-nested CV. Unlike the proposed approach, the number of genes retained cannot be computed by the algorithms. Hence, they are usually chosen by user experience [17, 37]. As an illustration, the number of genes retained is 50 for these algorithms, which is still larger than the average number of retained genes computed using the proposed RPLS-RFE approach.

In closing, we remark that the proposed method is applicable to other diseases and other forms of data, which potentially can shorten significantly the time to identify preliminarily critical genes and pathways associated with known or new diseases from such high throughout biological data. The shortlisted genes can then be further investigated through biological experiments.

V. CONCLUSION

A new gene identification approach based on RPLS with RFE and novel Brier and OP measures has been presented. The estimation variance is improved by incorporating RPLS with RFE and the Brier error measure using a two-nested CV. Second, novel Brier and OP-based measures are utilized to rank genes across CV subsamples and detect different GRNs from correlated genes, which consistently appear in the ranking list. In the analysis of three publicly available data sets including leukemia [9], colon [24], and prostate [25], experimental results showed that cancer-related genes and their associated GRNs can be preliminarily identified, while the classification performance is highly comparable with other conventional approaches. These findings suggest that the proposed approach may serve as a useful tool for preliminary identification of biological significant genes and their associated GRNs associated with a particular disease for further biological studies using gene microarray or other similar data.

APPENDIX A

DERIVATION OF LOGISTIC REGRESSION

In this section, the solution of maximizing the log-likelihood $\sum_{i=1}^{N'} \left[y_i \ln p_i(\alpha, \beta) + (1 - y_i) \ln (1 - p_i(\alpha, \beta)) \right]$ in (10) is derived. It can be determined by setting the partial derivative of the log likelihood to zero. This yields [52]:

$$\frac{\partial \ln(L)}{\partial \alpha} = Y - p(\alpha, \beta) = 0 \quad (A1)$$
$$\frac{\partial \ln(L)}{\partial \beta} = X^T_{\text{TRAIN}} (Y - p(\alpha, \beta)) = 0 \quad (A2)$$

where $p(\alpha, \beta)$ is the classification probability of the new estimate $\tilde{p}_i$ obtained in (25) is defined as:

$$p_i(\alpha, \beta) = \frac{1}{1 + \exp(-\alpha - \beta^T x_i)}$$

where $\alpha$ and $\beta$ are approximated solution, say obtained in previous iteration

$$\frac{\partial p_i}{\partial \alpha} = \tilde{p}_i (1 - \tilde{p}_i), \quad (A4)$$
$$\frac{\partial p_i}{\partial \beta_j} = \tilde{p}_i (1 - \tilde{p}_i) \gamma_{ij} \quad (A5)$$

and $\tilde{p}_i = p_i(\alpha, \beta)$. To simplify notation, we further define

$$\tilde{X} = [1_{N'}, \tilde{X}_{\text{TRAIN}}] \text{ and } \gamma = [\alpha, \beta^T]^T$$

Combining (A4) and (A5), the desired solution satisfies

$$(\tilde{X}^T W \tilde{X}) \gamma \approx \tilde{X}^T (Y - \tilde{p} + \tilde{W} \tilde{X} \gamma) \quad (A7)$$

where $W = \text{diag} \{\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_{N'}\}$, $\tilde{w}_i = \tilde{p}_i (1 - \tilde{p}_i)$, and $\tilde{p}_i = p_i(\bar{\alpha}, \bar{\beta})$. By solving (A6) and using the solution as the new estimate $\gamma = [\bar{\alpha}, \bar{\beta}]^T$ repeatedly, one obtains the IRWLS algorithm for solving the regression coefficients.

APPENDIX B

DERIVATION OF PARTIAL LEAST SQUARES

In this section, the details of solving the optimization problem

$$\max \{ e^{(m)^T} u^{(m)} \| \omega^{(m)} \|^2_2 = 1, \| q^{(m)} \|^2_2 = 1 \}$$

in (25) is derived. By means of the Lagrange multiplier method, $\omega^{(m)}$ and $q^{(m)}$ can be solved as an eigenvalue problem. To this end, we form the Lagrangian function as

$$L = e^{(m)^T} u^{(m)} - \lambda_1 \left( \| \omega^{(m)} \|^2_2 - 1 \right) - \lambda_2 \left( \| q^{(m)} \|^2_2 - 1 \right) \quad (B1)$$
where $\lambda_1$ and $\lambda_2$ are the Lagrange multipliers. The optimal solution is obtained by setting the gradient of $L$ with respect to $\omega^{(m)}$ and $q^{(m)}$ to zero. This yields $X^{(m)T}Z^{(m)}q^{(m)} = 2\lambda_1^{(m)}\omega^{(m)}$ and $Z^{(m)T}X^{(m)}\omega^{(m)} = 2\lambda_2^{(m)}q^{(m)}$. Further eliminating $q^{(m)}$ gives $X^{(m)T}Z^{(m)}Z^{(m)T}X^{(m)}\omega^{(m)} = \lambda_1^{(m)}\lambda_2^{(m)}\omega^{(m)}$. Hence, $\omega^{(m)}$ is the eigenvector of $C^{(m)} = C_{xx}^{(m)}C_{xz}^{(m)} = X^{(m)T}Z^{(m)}Z^{(m)T}X^{(m)}$ and

$$q^{(m)} = \frac{\sqrt{\omega^{(m)T}X^{(m)T}Z^{(m)}Z^{(m)T}X^{(m)}\omega^{(m)}}}{\sqrt{\omega^{(m)T}X^{(m)T}Z^{(m)}Z^{(m)T}X^{(m)}\omega^{(m)}}},$$

$\omega = \omega^{(m)} + \frac{\lambda_1^{(m)}}{\lambda_2^{(m)} - \lambda_1^{(m)}}q^{(m)}$.

Let $X^{(m)T}Z^{(m)} = \sum_{i=1}^{R} \sigma_i^{(m)} v_i^{(m)} v_i^{(m)T} = U^{(m)} \Lambda^{(m)} V^{(m)T}$ be the singular value decomposition of $X^{(m)T}Z^{(m)}$, where $R$ is the number of basis of $X^{(m)T}Z^{(m)}$. It can be seen that $\omega^{(m)}$ is the first singular vector of $X^{(m)T}Z^{(m)}$ while $q^{(m)}$ is the right singular vector of $X^{(m)T}Z^{(m)}$ with $\sigma_1^{(m)} = 2\lambda_1^{(m)}$. After $t^{(m)}$ is determined from $\omega^{(m)}$ using (24), we can determine $p^{(m)}$ from the rank-one approximation to $X^{(m)}$ as

$$X^{(m)} = t^{(m)} p^{(m)T} + E^{(m)}.$$  

If $E^{(m)}$ is normal distributed, then $p^{(m)}$ can be determined as

$$\min_{p^{(m)}} \left\| W_2^{(m)} X^{(m)} - W_1^{(m)} t^{(m)} p^{(m)T} \right\|^2_2$$

and its solution is $p^{(m)} = (X^{(m)T}W_1^{(m)})/(t^{(m)T}t^{(m)})$. Note, from (13), we can see that $Z$ is weighted by $X^{(m)T}W_1^{(m)}$ and therefore both $Z$ and $X$ should be weighted by $W_1^{(m)}$. Similarly, $e^{(m)}$ can be determined as

$$\min_{e^{(m)}} \left\| W_2^{(m)} Z^{(m)} - W_1^{(m)} t^{(m)} e^{(m)T} \right\|^2_2$$

and its solution is $e^{(m)} = Z^{(m)T}W_1^{(m)T}t^{(m)T}$. Since $Z^{(m)T}W_1^{(m)} = (W_1^{(m)T}Z - W_1^{(m)T} \sum_{i=1}^{m-1} t_i c_i^{T}) W_1^{(m)T} t^{(m)} = \hat{Z}^{T} W_1^{(m)}$, we have $C^{(m)} = (\Gamma W D^{(-1)})^{T} Z$ where $\Gamma = [t^{(1)}, \ldots, t^{(M)}]$ and $M$ is the number of PC retained. The process can be repeated by removing the contributions of these components from $X^{(m)}$ and $Z^{(m)}$ as

$$X^{(m+1)} = X^{(m)} - t^{(m)} p^{(m)T}, \quad Z^{(m+1)} = Z^{(m)} - t^{(m)} e^{(m)T}.$$  

The final regression vector $\gamma_{PLS}$ is obtained as

$$\gamma_{PLS} = \Phi (P^{T}\Phi)^{+} C^{T},$$

with $\Phi = [\omega^{(1)}, \ldots, \omega^{(M)}].$
S. C. Chan (S’87–M’92) received the B.Sc. (Eng.) and Ph.D. degrees from the University of Hong Kong, Pokfulam, Hong Kong, in 1986 and 1992, respectively.

In 1992, he joined City Polytechnic of Hong Kong, Kowloon, Hong Kong, as an Assistant Lecturer and later as a University Lecturer. Since 1994, he has been with the Department of Electrical and Electronic Engineering, the University of Hong Kong, and is now a Professor. He was a Visiting Researcher in Microsoft Corporation, Redmond, WA, Microsoft, Beijing, China, University of Texas at Arlington, Arlington, and Nanyang Technological University, Singapore. His research interests include fast transform algorithms, filter design and realization, multirate and biomedical signal processing, communications and array signal processing, high-speed AD converter architecture, bioinformatics, and image-based rendering.

Dr. Chan is currently a Member of the Digital Signal Processing Technical Committee of the IEEE Circuits and Systems Society and the Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I and Journal of signal processing systems. He was the Chairman of the IEEE Hong Kong Section and a member of the Board of Governors of the IEEE Circuits and Systems Society. He was a Visiting Scholar at the National University of Singapore, the University of California, Berkeley, and the University of Tokyo, Japan. He is currently working as an Associate Professor in Microsoft Corporation, Redmond, WA, Microsoft, Beijing, China, University of Texas at Arlington, Arlington, and Nanyang Technological University, Singapore. His research interests include fast transform algorithms, filter design and realization, multirate and biomedical signal processing, communications and array signal processing, high-speed AD converter architecture, bioinformatics, and image-based rendering.

H. C. Wu (S’10) received the B.Eng. degree in electronic engineering from the University of New South Wales, Sydney, Australia, in 2008, and the B.Sc. degree in electrical and electronic engineering from the University of Hong Kong, Pokfulam, Hong Kong, in 2009, where he is currently working toward the Ph.D. degree in the Department of Electrical and Electronic Engineering.

His main research interests are in pattern recognition, digital signal processing, and bioinformatics.

K. M. Tsui received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The University of Hong Kong, Pokfulam, Hong Kong, in 2001, 2004, and 2008, respectively. He is currently working as a Postdoctoral Fellow in the Department of Electrical and Electronic Engineering at the same university. His main research interests are in array signal processing, high-speed AD converter architecture, biomedical signal processing, digital signal processing, multirate filter bank and wavelet design, and digital filter design, realization, and application.