Robust cross-race gene expression analysis

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/ICASSP.2009.4959631">http://dx.doi.org/10.1109/ICASSP.2009.4959631</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Sun Mar 10 00:42:08 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/58098">http://hdl.handle.net/1721.1/58098</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
ROBUST CROSS-RACE GENE EXPRESSION ANALYSIS

Hsun-Hsien Chang and Marco F. Ramoni

Children’s Hospital Informatics Program, Harvard Medical School, Boston, MA
Division of Health Sciences and Technology, Harvard-MIT, Boston, MA

ABSTRACT

This paper develops a Bayesian network (BN) predictor to profile cross-race gene expression data. Cross-race studies face more data variability than single-lab studies. Our design handles this problem by using the BN framework. In addition, unlike existing methods that unrealistically assume independent genes, our BN approach can capture the dependencies among genes. Existing BN algorithms in biomedicine applications quantify data, leading to information loss; we adopt linear Gaussian model to keep the data intact, so our resulting model is more reliable. The application of our BN predictor to a lung adenocarcinoma study shows high prediction accuracy, and performance evaluation demonstrates our gene signature agreeable with those reported in the literature. Our tool has a promising potential in finding disease biomarkers common to multiple races.

Index Terms— gene expression, Bayesian networks, transcriptional diagnosis, cross-race studies.

1. INTRODUCTION

Comparative analysis of gene expression levels between multiple tissue states makes transcriptional diagnosis feasible [1]. The analysis starts with identifying a signature of gene transcripts that differentially express across tissue conditions, and then constructs a tissue classifier using the signature. A decade ago, expression studies were conducted by single-lab analysis; i.e., the training data and the independent testing data were collected from the same research lab. Along with the advancement in microarray technology, gene expression profiling becomes a widely accepted technique in many molecular biology labs. As such, researchers can test the generalizability of signatures beyond lab boundaries. Cross-lab studies seek biomarkers by the data acquired from one institute, and then test the predictive performance of the signature using the data obtained from another institute. Cross-lab expression analysis is more challenging because the data collected in different labs has more variability, induced by nonuniform experimental protocols such as RNA sample preparation and microarray operations [2].

Cross-race studies are a new application area of gene expression profiling. The task of cross-race studies is to look for disease biomarkers common to multiple races. Besides having the same sources of data variability as cross-lab studies, cross-race data experience other variability due to distinct patient populations. Due to non-identical living environment, the genes in different races express diversely, so the expression levels of the same set of biomarkers vary. To handle the data variability arising from multiple data sources, we need a robust analysis tool, which is the goal of this paper.

Most of existing works were designed in the era of single-lab studies. Popular techniques can be categorized into data-driven and model-driven approaches. Data-driven methods, such as fold change [3], t statistic [4], or signal to noise ratio [5], rank all the genes based on the statistical measures of their expression levels. Model-driven methods describe the microarray data by probabilistic models and rank the genes based on a measure quantifying the model difference between tissue conditions [6]. The genes with measures exceeding an empirically determined threshold assemble a signature. Data-driven approaches are easily vulnerable to any data variability, so we opt for the model-driven approach to process multi-race data. Unlike current model based schemes, our design needs a more sophisticated model which is robust to cross-race data variability.

To avoid the predictive performance deteriorated by data variability, we consider two aspects in the design. First, we adopt a probabilistic model to describe the expression data and to regularize decision making. Second, existing methods assume that genes are independent, contradicting to the reality that genes interact directly or indirectly in biological processes. We propose to incorporate our classifier design with a more realistic network model capable of describing these dependencies. Among various design paradigms, we choose the Bayesian network (BN) framework, which is a probabilistic graph model, to meet our needs.

Besides handling data variability and capturing gene dependencies, our BN approach has the following features.

- In gene expression data, the phenotype is a discrete variable taking category numbers and the genes are continuous variables with expression levels ranging from zero to infinity. Existing BN based methods [7] in biomedicine applications quantize variables to infer the optimal BN for the data, but quantization results in information loss. In contrast, we keep the data intact by adopting the linear Gaussian model to explore the dependencies among genes, yielding a more genuine BN model.

- Our signature search is capable of eliminating collinearly expressed genes. When a gene expresses collinearly with a biomarker, existing methods tend to include it in the signature. We avoid this problem by evaluating the likelihood of the gene’s dependence on the phenotype or on another gene. If the gene is most likely dependent on the phenotype, it is a biomarker, and our BN model depicts it as modulated by the phenotype. For example, Figure 1 presents a BN describing a data set of six genes. Genes 1, 2, 3 are the biomarkers and modulated by the phenotype; genes 5 and 6 are not biomarkers but are collinear with gene 3, so the BN describes them as modulated by gene 3.

- Our BN based approach is threshold free to determine biomarkers. After computing scores of genes, existing works have to cut off the list by assigning a threshold. Unlike these methods that require subjective thresholds, our BN approach has determined the signature genes once the optimal network is
Fig. 1. Illustration of a Bayesian network.

learned from the data. The signature genes for sample classification are the genes modulated by the phenotype. Other genes not modulated by the phenotype do not play a role in classification, so they can be discarded. With reference to Figure 1, genes 1, 2, 3 assemble a signature for tissue classification; genes 4, 5, 6 can be discarded because of their irrelevance to the classification task.

2. METHODS

The BN framework for gene expression analysis consists of two steps: learn the optimal BN characterizing the given data and develop the corresponding classification scheme on testing samples. This section starts with the algorithm for learning optimal BN with linear Gaussian model, and then describes how to make prediction in our model.

2.1. Learning Bayesian Network with Linear Gaussian Model

Let \( Y_1, Y_2, \cdots, Y_C \) be Gaussian random variables representing the expression levels of genes, and \( C \) be a binomial random variable characterizing two tissue conditions. We use uppercase to denote random variables and lowercase to denote their values. Given the gene expression data \( D = \{ y_1, \cdots, y_C, c \} \), the task is to find the best BN model from a set of candidate models \( M = \{ M_1, \cdots, M_K \} \) or, equivalently, searching for the largest posterior probability \( p(\theta_k|D) \). Applying Bayes’ theorem to \( p(M_k|D) \) results in

\[
p(M_k|D) \propto p(\theta_k)p(D|M_k),
\]

where \( p(\theta_k) \) is the prior probability of each model and \( p(D|M_k) \) is the marginal likelihood. The computation of \( p(D|M_k) \) is to average out \( \theta_k \) from the likelihood function \( p(D|\theta_k) \), where \( \Theta_k \) is the random vector parameterizing the distribution of \( Y_1, Y_2, \cdots, Y_C \) conditional on \( M_k \). We can exploit the local Markov properties encoded by the network \( M_k \) to rewrite the joint probability \( p(D|\theta_k) \) as

\[
p(D|\theta_k) = p(c|pa(c), \theta_{kc}) \prod_{g=1}^G p(y_g|pa(y_g), \theta_{kg}),
\]

where \( \text{pa}(y_g) \) denotes the values of the parents \( \text{Pa}(Y_g) \) of \( Y_g \), and \( \theta_{kg} \) is the subset of parameters used to describe the dependence of \( Y_g \) on its parents.

In this paper, we model a gene \( Y_g \) to be dependent on either the phenotype \( C \) or another single gene \( Y_a \), and the phenotype \( C \) is a root in the network without parents. We further can assume the \( J \) samples in the database are independent. The likelihood function becomes

\[
p(D|\theta_k) = \left[ \prod_{j=1}^J p(c_j|\theta_{kc}) \right] \times \left[ \prod_{j=1}^J \prod_{g=1}^G p(y_{gj}|\text{pa}(y_{gj}), \theta_{kg}) \right],
\]

where the subscripts \( j \) indicate the \( j \)th sample. The first term can be estimated by the sample frequencies: \( J_A (1 - J_A)^{J-A} \), where \( J_A \) and \( J \) are the number and the frequency parameter of the samples occurred in tissue condition \( A \), respectively. The second term is computed by the linear Gaussian model [8]. When the parent of \( Y_g \) is another gene \( Y_a \), i.e., \( \text{Pa}(Y_g) = Y_a \), the conditional mean is a first order linear regression

\[
\mu_g = \beta_{g0} + \beta_{g1} y_a.
\]

When \( \text{Pa}(Y_g) = C \), the conditional mean of \( Y_g \) is parameterized by \( c \):

\[
\mu_g = \beta_{g0}(c).
\]

It follows that

\[
p(y_{gj}|\text{pa}(y_{gj}), \theta_{kg}) = \frac{(\tau_g/2\pi)^{1/2}}{\sqrt{\left| \Sigma_g \right|}} \exp \left( -\frac{1}{2\tau_g} \left( y_{gj} - \mu_{g,j} \right)^T \left( y_{gj} - \mu_{g,j} \right) \right),
\]

where \( \mu_{g,j} \) denotes the conditional mean of \( Y_g \) in sample \( j \), and the vector \( \theta_{kg} \) denotes the set of parameters \( \gamma_{g}, \beta_{g0}, \beta_{g1} \) in model \( M_k \).

It is more convenient to adopt matrix notation to write the likelihood function in a compact form. We use the vector \( c = [c_1, \cdots, c_J]^T \) to denote the sample phenotypes, the vector \( y_g = [y_{g1}, \cdots, y_{gJ}]^T \) to stack the observations of \( Y_g \), the vector \( \beta_g = [\beta_{g0}, \beta_{g1}]^T \) to collect the regression coefficients, and the matrix

\[
X_g = \begin{bmatrix} 1 & \text{pa}(y_{g1}) \\ \vdots & \vdots \\ 1 & \text{pa}(y_{gJ}) \end{bmatrix}
\]

to denote the expression values of parents of \( y_g \). When \( \text{Pa}(Y_g) = C \), \( \beta_g = [\beta_{g0}] \) and \( X_g = 1 \). It follows that the second term in the likelihood function becomes

\[
\prod_{g=1}^G \frac{(\tau_g/2\pi)^{1/2}}{\sqrt{\left| \Sigma_g \right|}} \exp \left( -\frac{1}{2\tau_g} \left( y_g - X_g \beta_g \right)^T(y_g - X_g \beta_g) \right).
\]

To compute the marginal likelihood, we need to learn the distributions of \( \tau_g \) and \( \beta_g \). The standard conjugate prior for \( \tau_g \) is a Gamma distribution

\[
\tau_g \sim \Gamma(\alpha_{g1}, \alpha_{g2}), \quad p(\tau_g) = \frac{1}{\alpha_{g2}^{\alpha_{g1}}} \Gamma(\alpha_{g1}) \tau_g^{\alpha_{g1}-1} e^{-\alpha_{g2}/\tau_g}
\]

where \( \alpha_{g1} = \frac{\nu_{g0}}{2} \) and \( \alpha_{g2} = \frac{1}{\nu_{g0}\sigma_{g0}^2} \) are characterized by hyper-parameters \( \nu_{g0}, \sigma_{g0}^2 \). The marginal expectation of \( \tau_g \) is

\[
E(\tau_g) = \alpha_{g1}/\alpha_{g2} = \frac{1}{\sigma_{g0}^2}
\]

and

\[
E(1/\tau_g) = \frac{1}{(\alpha_{g1} - 1)\alpha_{g2}} = \frac{\nu_{g0}\sigma_{g0}^2}{\nu_{g0} - 2}
\]

is the prior expectation of the population variance. Because \( E(1/\tau_g) \) is similar to the estimate of the variance in a sample of size \( \nu_{g0}, \sigma_{g0}^2 \) is the prior population variance, based on \( \nu_{g0} \) cases seen in the past.
Conditional on $\tau_g$, the prior density of the parameter vector $\beta_g$ is supposed to be multivariate Gaussian:

$$\beta_g|\tau_g \sim N \left( b_{g0}, (\tau_g R_{g0})^{-1} \right)$$  \hspace{1cm} (12)

where $b_{g0} = E(\beta_g|\tau_g)$, $R_{g0}$ is the identity matrix so that the regression coefficients are a priori independent, conditional on $\tau_g$.

It can be shown that the marginal likelihood is

$$p(\{y_1, \ldots, y_G, c\}|M_k) = \frac{1}{(2\pi)^{n_g}/2} \prod_{g=1}^{G} \frac{1}{\nu_{g0}^2} \frac{1}{\sigma_{g0}^2} \times$$

$$\left( \frac{\nu_{g0}}{2} \Gamma(\nu_{g0}/2) \right)^{(n_g+1)/2} \frac{1}{\sigma_{g0}^2/(\nu_{g0}-2)}$$

where the parameters are specified by the following rules:

$$\alpha_{g1n} = \nu_{g0}/2 + J/2$$ \hspace{1cm} (14)

$$R_{gn} = R_{g0} + X_g^T X_g$$ \hspace{1cm} (15)

$$b_{gn} = R_{gn}^{-1} (R_{g0} b_{g0} + X_g^T y_g)$$ \hspace{1cm} (16)

$$\frac{1}{\alpha_{g2n}} = (-b_{gn}^T R_{gn} b_{gn} + y_g^T y_g + b_{g0}^T R_{g0} b_{g0})/2 + 1/\sigma_{g2}$$ \hspace{1cm} (17)

$$\nu_{gn} = \nu_{g0} + J$$ \hspace{1cm} (18)

$$\sigma_{gn} = \sqrt{\nu_{g0}^2/\nu_{g0} - 2}$$ \hspace{1cm} (19)

The Bayesian estimates of the parameters are given by the posterior expectations:

$$E(\tau_g | y_g) = \alpha_{g1n} \alpha_{g2n} = 1/\sigma_{g2}$$ \hspace{1cm} (20)

$$E(\beta_g | y_g) = b_{gn}$$ \hspace{1cm} (21)

$$E(1/\tau_g | y_g) = \nu_{g0} \sigma_{g2}^2/(\nu_{g0} - 2)$$ \hspace{1cm} (22)

The selection of the best BN model $\hat{M}$ relies on the Bayes factor, $BF$. For arbitrary two candidate models $M_k$ and $M_h$, their Bayes factor is

$$BF_{kh} = \frac{p(M_k)p(D|M_k)}{p(M_h)p(D|M_h)}$$ \hspace{1cm} (23)

If $BF_{kh} \geq 1$, we choose model $\hat{M} = M_k$; otherwise, $\hat{M} = M_h$.

The criterion for performance evaluation is the area under receiver operating characteristic (AUROC) curve. The quantity of AUROC ranges from 0 to 1; the higher the AUROC is, the better the predictor has. The fitted validation, i.e., predicting the training set itself, yields 100% AUROC. The prediction on the independent Taiwanese data produces 95% AUROC.

Besides the performance evaluation by AUROC, we examine the biological quality of the 12 biomarkers. Table 1 summarizes the 12 signature genes and their functions revealed in the literature. Except HBA and SPINK1, the other 10 genes are discovered to be related to lung cancer or a subtype of lung cancer, confirming good quality of our method. We briefly discuss the biomarkers in the following:

- **FAM20B**, **MUC5B**, **SFTP C**, and **XAGE1** have been reported as biomarkers to lung adenocarcinoma.
- **KRT6**, **KRT16**, and **MAGEA2** are biomarkers of squamous carcinoma; since adenocarcinoma and squamous carcinoma are both the subtypes of non-small-cell lung cancer, these 3 biomarkers explain that there is similarity between the two subtypes of lung cancer.
- **CXCL13** and **SERPINB5** have been known as biomarkers of lung cancer, so it is not surprising that they are predictive on adenocarcinoma.
- **MARCO** expresses when the lung is exposed to smoke, although it is not directly related to lung cancer. It is common that smokers have higher probability of getting lung cancer,

---

**3. RESULTS AND DISCUSSION**

We apply our method to studying molecular biomarkers of lung adenocarcinoma. The training data includes 107 subjects from the Lombardy region in Italy [9], which is publicly available on Gene Expression Omnibus (GEO) with accession number GSE10072; there are 49 controls and 58 cases. The testing data includes 63 subjects collected in Taiwan [10], which consists of 31 controls and 32 cases and whose GEO accession number is GSE7670. The gene expression experiments were carried out by Affymetrix HG-U133A, which is equipped with 22,283 probes. Probe level analysis was performed using the Robust Multi-array Algorithm (RMA). The detailed protocols of sample preparation and the demographic information of patients were described in [9, 10].

After our algorithm learns the optimal BN, we trim away the genes not modulated by the phenotype, leading to the final predictive BN model shown in Figure 2. The rectangle node is the root indicating the phenotype and the 12 elliptic nodes are signature genes. We further evaluate the prediction performance using these 12 biomarkers. The criterion for performance evaluation is the area under receiver operating characteristic (AUROC) curve. The quantity of AUROC ranges from 0 to 1; the higher the AUROC is, the better the predictor has. The fitted validation, i.e., predicting the training set itself, yields 100% AUROC. The prediction on the independent Taiwanese data produces 95% AUROC.

Besides the performance evaluation by AUROC, we examine the biological quality of the 12 biomarkers. Table 1 summarizes the 12 signature genes and their functions revealed in the literature. Except HBA and SPINK1, the other 10 genes are discovered to be related to lung cancer or a subtype of lung cancer, confirming good quality of our method. We briefly discuss the biomarkers in the following:

- **FAM20B**, **MUC5B**, **SFTP C**, and **XAGE1** have been reported as biomarkers to lung adenocarcinoma.
- **KRT6**, **KRT16**, and **MAGEA2** are biomarkers of squamous carcinoma; since adenocarcinoma and squamous carcinoma are both the subtypes of non-small-cell lung cancer, these 3 biomarkers explain that there is similarity between the two subtypes of lung cancer.
- **CXCL13** and **SERPINB5** have been known as biomarkers of lung cancer, so it is not surprising that they are predictive on adenocarcinoma.
- **MARCO** expresses when the lung is exposed to smoke, although it is not directly related to lung cancer. It is common that smokers have higher probability of getting lung cancer,
so MARCO is a reasonable biomarker for predicting adenocarcinoma.

- Although HBA and SPINK1 have not been reported for their association with any subtypes of lung cancer, our result suggests that it is worthwhile to study their biological function in lung cancer.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Function Reported in Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAM20B</td>
<td>abundant in lung and differentially expressed in lung adenocarcinoma [12]</td>
</tr>
<tr>
<td>HBA1/HBA2</td>
<td>n/a</td>
</tr>
<tr>
<td>KRT6A/B/C</td>
<td>biomarker of lung squamous cancer [13]</td>
</tr>
<tr>
<td>KRT16</td>
<td>biomarker of lung squamous cancer [14]</td>
</tr>
<tr>
<td>MAGEA2/2B</td>
<td>biomarker of lung squamous cancer [15]</td>
</tr>
<tr>
<td>MARCO</td>
<td>upregulated when exposed to lipopolysaccharides and smoke [16]</td>
</tr>
<tr>
<td>MUC5B</td>
<td>biomarker of lung adenocarcinoma [17]</td>
</tr>
<tr>
<td>SERPINB5</td>
<td>biomarker of lung cancer [18]</td>
</tr>
<tr>
<td>SFTPC</td>
<td>biomarker of lung adenocarcinoma [19]</td>
</tr>
<tr>
<td>SPINK1</td>
<td>n/a</td>
</tr>
<tr>
<td>XAGE1/1B/1C/1D/1E</td>
<td>biomarker of lung adenocarcinoma [20]</td>
</tr>
</tbody>
</table>

Table 1. The 12-gene signature for lung adenocarcinoma diagnosis.

4. CONCLUSIONS

This paper develops a gene expression analysis algorithm in the BN framework for cross-race studies. Unlike prior works, our development adopts linear Gaussian model and considers more realistic biology that genes are dependent through their molecular interactions. The application of our BN predictor to an international lung adenocarcinoma study demonstrates how the BN method solves the real world problem. The BN predictor obtains 12 biomarkers. The prediction on an independent data set using this 12-gene signature reaches 0.95 AUROC, showing good generalizability. The biological confirmation agrees our signature with the lung cancer genes in the literature. The proposed method will have a potential to perform clinical cross-race transcriptional diagnoses.

5. REFERENCES


