Integrated Analysis on the Gene Expression and Methylation Microarray Data and the Application on Biomarker Development of Ovarian Cancer
Xueming Zhang
Co-advisors: Hongbin Fang, PhD and James Li, PhD
Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University

Introduction
Ovarian Cancer is an important cancer among and the fifth leading cause of cancer death among women in the United States. The exact cause is still unknown. The majority of the cancer can be attributed to a growing number of somatic aberrations (Bast, 2009) and the rest may due to germline mutations (Pal 2005; Risch, 2006). The successful treatment for this cancer is currently unavailable. The standard treatment is aggressive surgery followed by chemotherapy, and the cancer recurs rate is about 25% with 6 months (Miller 2009). To better understand the molecular abnormalities that influence the pathophysiology and identify the potential targets for therapy. The Cancer Genome Atlas (TCGA) researchers studied a series of genomics and epigenomics data including expression, DNA copy number, microRNA expression, exome DNA sequencing. These comprehensive datasets greatly facilitated the analyses of the gene expression, methylations and genotypes and variations related to cancer development. Despite the analyses on gene expression or promoter methylation data individually for the cancer biomarkers, little is done on combining their information in the same model for an integrated analysis. DNA methylation has been thought as being an essential role of gene expression, usually negatively correlated with gene expression. Genomic data obtained from gene, transcript or methylation studies may have both common and unique information regarding the disease progress. In this paper, we propose a novel variable selection approach to find an optimal panel of biomarkers to detect cancer. The approach is using iterative lasso regressions accounting for the relationship between gene expressions and DNA methylations. The analysis of datasets including 136 health subject and 338 clinically annotated stages II to IV high-grade serous ovarian adenocarcinomas (HGS-OvCa) illustrates the proposed approach is effective. The result suggests that integrating more information from the comprehensive cancer genomic study data could significantly improve the efficiency of the biomarkers for disease diagnosis and prognosis.

Methods

Variable Selection with LASSO regression
Let Y denote true disease status of subject i (Yi = 1 if subject is diseased and Yi = 0 if non-diseased) for i = 1, 2, ..., n. Let Xi denote the q-dimensional vector of covariates associated with subject i and 0 be a q-dimensional vector of unknown coefficients. We consider the following logistic model,

\[
P(Y_i = 1 | x_i, \beta) = \frac{1}{1 + e^{-x_i' \beta}}
\]

where \( \beta = (\beta_1, \beta_2, ..., \beta_q) \) is a tuning parameter.

Then, the penalized maximum likelihood estimate (MLE) of \( \hat{\beta} \) is given by

\[
\hat{\beta} = \arg \max \{ \log \sum_{i=1}^{n} P(Y_i | x_i, \beta) - \frac{1}{2} \| \beta \|_1 \}
\]

where \( \lambda > 0 \) is a penalty parameter, and selected by minimizing the approximate generalized cross-validation (GCV) statistic. The GCV in this case is based on both sensitivity and specificity of the predictive model (Tibshirani, 1996).

A Novel Procedure for Variable Selection:
Suppose there a gene expressions and t DNA methylations
1. Using correlation analysis to determine the relationships between gene expressions and DNA methylations based on the data of healthy subjects, the methylation variables are divided into two parts, \( \{ \hat{m}_1, \hat{m}_2, ..., \hat{m}_s \} \) are relative the gene expressions \( \{ g_1, g_2, ..., g_s \} \) and \( \{ m_1, m_2, ..., m_t \} \) are not correlated to the gene expressions.
2. Let the \( (i + 1 - k) \) - dimensional \( x^{(i)} \) be consist of the gene expressions and methylations which are not correlated to the gene expressions, i.e.,

\[
x^{(i)} = \{ \hat{x}_1, \hat{g}_1, \hat{g}_2, ..., \hat{g}_s, \hat{m}_1, \hat{m}_2, ..., \hat{m}_t \}
\]

Using the lasso regression, the \( (i + v - k) \) variables are selected for, example,

\[
x^{(i)} = \{ \hat{x}_1, \hat{g}_1, \hat{g}_2, ..., \hat{g}_s, \hat{m}_1, \hat{m}_2, ..., \hat{m}_t \}
\]

3. Since there are \( (i - s) \) genes deleted in the Lasso regression, assume that there are \( j \) methylated genes correlated to those \( (s-v) \) genes

\[
m_1, m_2, ..., m_j \subset \{ m_1, m_2, ..., m_t \}
\]

Let \( (i + v + k) \) - dimensional \( x^{(i)} \) be consist of \( x^{(i)} \) and the j methylation,

\[
x^{(i)} = \{ \hat{x}_1, \hat{g}_1, \hat{g}_2, ..., \hat{g}_s, \hat{m}_1, \hat{m}_2, ..., \hat{m}_j, \hat{m}_{j+1}, ..., \hat{m}_t \}
\]

The further variable selection is done by the Lasso regression with \( x^{(i)} \) again.

4. The procedure will be done until all genes in the selected covariates are significant in the predict model.

After the variable selection with the iterative Lasso regression, the receiver operating characteristic (ROC) analysis is used to assess the performance of these LASSO regressions. The predicted values of disease status were calculated based on the coefficients obtained above. ROC curves and the areas under curve (AUC) were computed by the R package “pROC” using the predicted disease status and true status of each dataset.

Software
All statistical analyses were performed by R version 3.0.3 on a Windows operation system.

Results
136 health subjects and 338 ovarian cancer patients. For each subject, 11,864 genes and 23,678 methylations are measured. With a significant level of p-value <0.005, 130 genes and 42 methylations are selected as candidate predictors for disease status by t-test.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Lambda</th>
<th>AUC</th>
<th>Genes</th>
<th>Methylations</th>
<th>Total variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>130 genes</td>
<td>0.0180556</td>
<td>0.8368</td>
<td>44</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>42 methylations</td>
<td>0.0080586</td>
<td>0.8102</td>
<td>29</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>All methylations (23,678) + all genes (11,864) (ignoring their relationships)</td>
<td>0.0164516</td>
<td>0.7216</td>
<td>9</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>

Using the proposed approach

0.0136584 | 0.8750 | 32 | 21 | 53 |

Conclusion
To find an optimal panel of biomarkers, a novel variable selection approach is proposed based on iterative lasso regressions accounting for the relationship between gene expressions and DNA methylations. The receiver operating characteristic (ROC) analysis is used to assess the performance of these LASSO regressions. The result suggests that integrating more information from the comprehensive cancer genomic study data could significantly improve the efficiency of the biomarkers for disease diagnosis and prognosis.

References