Abstract

A breast cancer gene signature was constructed from expression profiles of peripheral blood cells using a novel method called LPLS. The potential of using gene expression for early diagnosis has been shown previously using internal double-cross validation. In this study it was shown that a gene signature from one expression study performs well in predicting cancer status in an independent sample set produced on the same expression platform. It is also shown that exploiting background information on gene relations may improve classifier performance.

Objectives

- Prediction of breast cancer status using gene expression profiles from peripheral blood cells
- Verification of gene signature on independent test samples
- Verification of the potential of the new LPLS method which exploits background information on gene-gene relations for classifier construction

Methods

High-density oligo-nucleotide arrays from Applied Biosystems were used to track the changes in genetic activity in peripheral blood. The expression profiling was done in parallel (same protocols, but in different batches) for samples belonging to two different studies.

The “DiaGenic” data from the first study (130 samples) were produced for the development of a prototype for early breast cancer diagnosis. Results from the analysis of these data have been reported previously [1].

The “MDG” data from the second study (50 samples) were part of a project aiming at increasing the knowledge about the biology underlying mammographic density and increased risk of breast cancer development [2]. Both data sets were pre-processed and batch adjusted in the same manner. The DiaGenic set was used as training set in the study reported here (STEP 1), whereas the MDG set was used as test data to verify the previous results from the DiaGenic data (STEP 2).

A novel method called LPLS-regression (Sæbø et al. [3]) was used for classifier construction. The LPLS utilizes background knowledge on the variables in the process of predictor construction in an attempt to reduce the influence from false positive genes and random noise. The predictor was trained on a subset of 486 genes from the DiaGenic dataset under the influence of information on inter-gene dependencies extracted from the KEGG-data base [4]. The 486 genes were selected due to their presence in the background information network. The reaction network information from KEGG was translated to a so-called modularity matrix expressing the degree of relation between genes. The performance of the LPLS classifier was compared to a regular PLS classifier which does not utilize background information on gene relations.

Results

The LPLS classifier was compared to a regular PLS classifier which utilizes background knowledge on the variables in the process of predictor construction in an attempt to reduce the influence from false positive genes and random noise. The predictor was trained on a subset of 486 genes from the DiaGenic dataset under the influence of information on inter-gene dependencies extracted from the KEGG-data base [4]. The 486 genes were selected due to their presence in the background information network. The reaction network information from KEGG was translated to a so-called modularity matrix expressing the degree of relation between genes. The performance of the LPLS classifier was compared to a regular PLS classifier which does not utilize background information on gene relations.

Table 1: Classifier performance on test data

<table>
<thead>
<tr>
<th></th>
<th>LPLS</th>
<th>PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>AUC</td>
<td>0.83</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Conclusions

The LPLS gene signature from the DiaGenic data performs well in predicting the cancer status of the 50 MDG samples. Even if the signature only used 486 probes, the accuracy was as high as 0.76. It is reasonable to expect improved accuracy if the number of probes with background information is increased. The test results verify the indications found by Sæbø et al. [2] that using background information on gene relations (LPLS vs PLS) may improve classifier performance.

References