Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers

Jonghwan Choi¹, Sanghyun Park², Youngmi Yoon³ and Jaegyoon Ahn¹,*

¹Department of Computer Science and Engineering, Incheon National University, Incheon, The Republic of Korea, ²Department of Computer Science, Yonsei University, Seoul, The Republic of Korea and ³Department of Computer Engineering, Gachon University, Seongnam-si, Gyeonggi-do, The Republic of Korea

*To whom correspondence should be addressed.

Abstract

Motivation: Identification of genes that can be used to predict prognosis in patients with cancer is important in that it can lead to improved therapy, and can also promote our understanding of tumor progression on the molecular level. One of the common but fundamental problems that render identification of prognostic genes and prediction of cancer outcomes difficult is the heterogeneity of patient samples.

Results: To reduce the effect of sample heterogeneity, we clustered data samples using K-means algorithm and applied modified PageRank to functional interaction (FI) networks weighted using gene expression values of samples in each cluster. Hub genes among resulting prioritized genes were selected as biomarkers to predict the prognosis of samples. This process outperformed traditional feature selection methods as well as several network-based prognostic gene selection methods when applied to Random Forest. We were able to find many cluster-specific prognostic genes for each dataset. Functional study showed that distinct biological processes were enriched in each cluster, which seems to reflect different aspect of tumor progression or oncogenesis among distinct patient groups. Taken together, these results provide support for the hypothesis that our approach can effectively identify heterogeneous prognostic genes, and these are complementary to each other, improving prediction accuracy.

Availability and implementation: https://github.com/mathcom/CPR

Contact: jgahn@inu.ac.kr

Supplementary information: Supplementary data are available at Bioinformatics online.
identify prognostic genes in network form (Ren et al., 2016; Wu and Stein, 2012). We previously proposed a network-based method to identify gene networks for the accurate diagnosis of prostate cancer (Ahn et al., 2011); this approach exploits the differences of correlations of gene pairs in the biological networks between two patient groups, but does not show good performance in application to prognosis data. Other approaches have been proposed for improving prediction of cancer outcomes (Roy et al., 2014; Winter et al., 2012). These methods identify prognostic genes using gene expression measurements and several gene networks including PPI networks. Genes are ranked according to their prognostic relevance using both expression and network information using NetRank, which iteratively calculates gene scores in a similar way to PageRank, but additionally uses correlation between gene expression values and patient survival time (Winter et al., 2012), or t statistics or fold-change of gene expressions between patients with good and poor prognosis (Roy et al., 2014).

Those methods commonly suffer from the heterogeneity of samples from cancer patients (Polyak, 2011), which makes prediction difficult. Since the main known source of the heterogeneity is genomic instability (Burrell et al., 2013), poor prediction may result from difficulties in identifying prognostic genes or biomarkers that are specific to certain cancer patients. Breast cancer has several genetic subtypes that have disparate clinical response and tumor progression (Russnes et al., 2011), so there have been many studies to deal with heterogeneity of breast cancer. Even one type of breast cancer, TNBC (triple-negative breast cancer), has six subtypes that show different biological characteristics (Lehmann et al., 2011).

A common approach to overcome the cancer heterogeneity problems is an integrative analysis of various omics data. Szeto et al. analyzed genomic, transcriptomic and proteomic data of TNBC patients at various clinical time-points to see temporal heterogeneity of TNBC (Szeto et al., 2017). Wang et al. proposed the network approach that integrates heterogeneous omics data and demonstrated that resulting glioblastoma and breast cancer patient groups show different survival profiles (Wang et al., 2016).

In the present study, we first clustered samples to deal with heterogeneity, using principal components of whole genes. For each cluster, we gave weights to edges in the functional interaction (FI) networks using t statistics of gene expressions between samples with good and poor outcomes in the cluster, and applied a modified PageRank algorithm to prioritize prognostic genes. Hub genes among resulting prioritized genes were selected as biomarkers to guarantee stable prediction accuracy among independent datasets. Then, we used Random Forest (Breiman, 2001) to predict prognosis of patient samples from gene expressions of the biomarkers.

This proposed approach outperformed the co-expression network approaches and NetRank, as well as traditional feature selection methods such as Lasso for predicting the prognosis of patients with breast cancer. We confirmed that clustering of heterogeneous samples actually contributes to prediction accuracy. In fact, we were able to find many cluster-specific prognostic genes for each dataset and functional study showed that distinct biological processes or pathways were enriched in each cluster; that is, our approach can actually enable detection of functional modules reflecting characteristics of patient groups. These results together provide support for the hypothesis that our approach can effectively identify heterogeneous prognostic genes and that these are complementary to each other, improving prediction accuracy.

Table 1. Summary of gene expression datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>#Total samples</th>
<th>#Poor samples</th>
<th>#Good samples</th>
<th>#Total genes</th>
<th>Characteristic for label</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA</td>
<td>189</td>
<td>90</td>
<td>99</td>
<td>20 501</td>
<td>days_to_death</td>
<td>GDAC firehose</td>
</tr>
<tr>
<td>GSE3494</td>
<td>157</td>
<td>36</td>
<td>121</td>
<td>13 181</td>
<td>Disease-Specific Survival Time</td>
<td>Miller et al. (2005)</td>
</tr>
<tr>
<td>GSE4922</td>
<td>175</td>
<td>69</td>
<td>106</td>
<td>13 181</td>
<td>Disease Free Survival Time</td>
<td>Ishihara et al. (2006)</td>
</tr>
<tr>
<td>GSE7390</td>
<td>140</td>
<td>24</td>
<td>116</td>
<td>10 501</td>
<td>Time of Overall Survival</td>
<td>Desmedt et al. (2007)</td>
</tr>
<tr>
<td>GSE24450</td>
<td>121</td>
<td>25</td>
<td>96</td>
<td>25 106</td>
<td>10 years follow up time</td>
<td>Heikkinen et al. (2011)</td>
</tr>
<tr>
<td>NKI</td>
<td>120</td>
<td>47</td>
<td>73</td>
<td>10 703</td>
<td>TIMEsurvival</td>
<td>van de Vijver et al. (2002)</td>
</tr>
</tbody>
</table>

2 Materials and methods

2.1 Data preparation

We downloaded one high-throughput sequencing (HTS) data from Broad Institute GDAC Firehose (Bitgda, 2016) and five microarray (MA) datasets from Gene Expression Omnibus (GEO) (Barrett et al., 2013) and van de Vijver et al. We used RSEM information for 20 501 genes from the HTS data of 189 patients. Four breast cancer prognostic datasets, GSE3494 (Miller et al., 2005), GSE4922 (Ishihara et al., 2006), GSE7390 (Desmedt et al., 2007), GSE24450 (Heikkinen et al., 2011) and their clinical information were collected from GEO. We also used 120 samples of van de Vijver dataset (van de Vijver et al., 2002). For all data, each sample was labeled as good prognosis if the patient survived more than 10 years (5 years for HTS dataset and GSE24450), and labeled as poor prognosis if the patient did not survive more than 5 years. Table 1 shows a summary of the described datasets. For all data, gene expression values were normalized for each sample by z-scoring.

His network is derived from curated pathways, protein-protein interactions, gene coexpression, Gene Ontology (GO) annotations and text-mined protein interactions (Wu et al., 2010). We downloaded the network data from Reactome (Croft et al., 2014; Fabregat et al., 2016), and the number of interactions is 150 322.

2.2 Cancer gene identification

We first computed principal components of gene expressions as attributes, and applied K-means clustering to samples with k with maximal silhouette score. For each cluster of samples, we computed weights of edges in FI networks with t statistics of genes. Then we applied the modified PageRank. It ranks genes by ratios of two PageRank scores which are calculated using weighted and unweighted networks. The modified PageRank uses the following formula:

$$s^*_i = (1 - d)s^0_i + \sum_{j \in NE_i} w_{ij} s^*_{j-1}$$

where $s^*_i$ denotes a score of gene i after n iterations, d is a damping
factor (user parameter, $0 < d < 1$), $NE_i$ is a set of genes which are neighbors of gene $i$, and $wij$ is calculated using:

$$wij = \begin{cases} \frac{|t_i|}{\sum_{k \in NE_i} |t_k|} & \text{on weighted networks} \\ \frac{1}{|NE_i|} & \text{on un-weighted networks,} \end{cases}$$

where $|NE_i|$ is the size of $NE_i$. A damping factor $d$ adjusts an influence of the FI network information. As a damping factor increases, an influence of the network also increases. The default value of $d$ is 0.7, which is found to be optimal through iterative experiments.

$t_i$ is $t$ statistics to test whether the means of expressions of gene $i$ are different between samples with poor and good outcomes. The initial scores $s_i$ have $s_i = G_{jj}$, where $G_{jj}$ is the number of genes, and thus the sum of scores is always 1. We decided that a score is converged if its change is less than 0.005, and observed the convergence after five iterations. So we computed average genes ranks to integrate gene sets from clusters, after five iterations. We can consider high-ranked genes as driver genes that are responsible for different prognoses.

The high-ranked genes do not all necessarily contribute to prediction of prognosis, because they may not show differences in expressions between poor and good outcomes, as effector genes do. Thus, we selected cut (user parameter, $50 \leq \text{cut} \leq 150$) hub genes whose degrees were in the top 2% (in our FIs, the degree of the top 2% is 185) as biomarkers. We found that 2% is optimal percentage rate through iterative experiments. We converted expression values of biomarkers to rank values for each sample to avoid test set bias (Patil et al., 2015), and applied Random Forest to the converted training and test data. We selected Random Forest since it showed the best result among widely used classification algorithms (Supplementary Fig. S1). The entire pipeline is provided in Figure 1 and implemented in Python with scikit-learn module (Pedregosa et al., 2011).

3 Results

3.1 Effect of sample clustering

The first step of our approach was clustering of samples, as they show different gene expression profiles. We were able to confirm that samples are roughly divided into two or three groups in PCA plots (Fig. 2). Note that those clusters did not show strong correlation with estrogen receptor (ER) status. Supplementary Table S1(a) shows that cluster1 and cluster2 of GSE4922, and cluster1 and cluster2 of GSE7390 have high ER$^+$ sample ratios. However, clusters with many ER$^-$ samples also had many ER$^+$ samples, which means that clustering does not necessarily reflect ER status. In fact, we found many ER$^+$ samples also had poor outcome, although generally ER$^-$ samples show poor outcome. As well as ER, PR status was not related with outcome (Supplementary Table S1b). We will show that the sample clustering actually contributes prediction accuracy and identification of driver genes, in subsequent chapters (Figs 3–5).

3.2 Performance of prognosis prediction

We performed 10-fold cross-validation and independent tests for different sets of two user parameters $\text{cut}$ and $d$. Complete results are provided in Supplementary Table S2.
While performing independent tests, we found that datasets showed very different distributions of gene expression, as shown in Figure 3. That is, training and test data were not compatible when we applied Random Forest and other classification algorithms. We focused on the identification of reproducible biomarkers that can be used for different independent datasets. Therefore, we designed independent tests as 10-fold cross-validation of test data with the gene set identified using training data. We used one dataset as training data and the other two as test datasets for independent tests. k in K-means clustering was set to 2 or 3 which maximized silhouetted scores. Optimal parameters are those with the best area under curve (AUC) in cross-validations (Supplementary Table S3).

We named our approach CPR (Clustering and PageRank), and compared it with CPR without clustering (hereafter, NCPR), Lasso (Sohn et al., 2009), NetRank (Roy et al., 2014), Wu and Stein (2012) and WGCNA (Langfelder and Horvath, 2008). Figure 4 shows ROC curve with BRCA as training data. We can see that our classification accuracy in terms of AUC is higher than others (0.001–0.094), except BRCA (training)-GSE4922 (test) and BRCA (training)-GSE24450 (test) case. CPR is generally superior to NCPR, which shows the effect of sample clustering. ROC curves from other datasets as training data are provided in Supplementary Figure S2, and it also shows CPR outperforms other approaches in general. The numbers of genes used for all methods and datasets, and the detailed information of genes using CPR is provided in Supplementary Tables S4 and S5, respectively.

We also investigated the minimal sample (good + poor) size that guarantees stable performance. Supplementary Figure S3 shows AUCs varying training sample sizes for all datasets, and we can check that the performance is stable if the sample size is greater than 80.

3.3 Analysis of oncogene inclusion ratio and reproducibility

Next, we counted known breast cancer oncogenes and tumor suppressor genes (TSG) collected from COSMIC (Forbes et al., 2017) and calculated the p-value using hypergeometric test for each dataset. Figure 5 shows that our genes had significantly low p-values overall, which means that the genes we identified have higher

![Fig. 3. PCA 3D scatter plot. PCA plots using three highest PCs for all datasets](image)

![Fig. 4. ROC curves. CPR is compared with NCPR (CPR without clustering), NetRank, Lasso, WGCNA and Wu & Stein, when using BRCA as training data. Complete ROC curves are provided in Supplementary Figure S2. Parameters and number of genes used are shown in Supplementary Tables S4 and S5, respectively](image)
probability of including breast cancer oncogenes or TSGs. Note that CPR shows better results than NCPR, which shows another effect of sample clustering. Complete contingency tables are provided in Supplementary Table S6.

We also calculated the ratio $r$ of overlapping prognostic genes among different datasets by following formula:

$$r = \frac{\text{(number of genes included in at least three data)}}{\text{(number of a union of the genes)}}$$

We observed that 37.8% of identified prognostic genes were overlapped at least three among different datasets in Figure 6, while 28.6%, 12.6%, 0.6% and 58.8% of genes were overlapped for Wu & Stein, WGCNA, Lasso and NetRank, respectively. NetRank has higher reproducibility, but identified genes are small number of extreme hub genes, and did not show good prediction accuracy.

The 37.8% genes that we identified include two famous oncogenes ERBB2 and TP53, and two tumor suppressor genes BRCA1 and CDH1. AKT1 was found only in GSE4922, CTCF was found in GSE4922 and NKI, and PIK3CA and CCND1 were found in both BRCA and GSE4922 (Fig. 7a–f). The PI3K-AKT1 signaling pathway has been studied as a clinical target for the breast cancer. Although AKT1 and PIK3CA1 are closely located on the pathway, PIK3CA1 may promote cancer through AKT-independent pathway (Vasudevan et al., 2009) and it has been also reported that their mutations have distinct effects on sensitivity to targeted pathway inhibitors in breast cancer model (Beaver et al., 2013).

### 3.4 Functional analysis of the gene module

We were able to identify many common biomarkers on both clusters, but there were also many cluster-specific biomarkers, as shown in Figure 7(a–f). Famous oncogenes and TSGs including BRCA1, TP53, ERBB2, CDH1, CTCF and PIK3CA were in common set, but also many oncogenes and TSGs were able to be found in cluster-specific sets. These results tell us that patient samples in different cluster can have different tumor growth patterns or tumorigenesis factors.

For more details, we performed functional analysis on biomarkers of clusters 0 and 1 of BRCA data. A biomarker belongs to cluster $x$ if its rank computed by our PageRank in cluster $x$ is (i) less than 1000 or (ii) less than ranks in all other clusters. The biomarkers of cluster 0 and 1, respectively, were test using DAVID (Huang da et al., 2009a,b) with default settings. Complete list of enriched GO terms (biological process) is provided in Supplementary Table S7. Among them, we selected some interesting GO terms ($\text{P-value } < 0.05$) enriched in each cluster exclusively and visualized them using Cytoscape (Shannon et al., 2003) in Figure 8.

We can observe many cluster-specific GO terms enriched, but we will first focus on the ‘response to leptin’ of cluster 0 and ‘integrin-mediated signaling pathway’ of cluster 1. Leptin is known to enhance breast cancer growth and progression (Andó et al., 2014), and it contributes metastasis in ER+ breast cancer cell (Strong et al., 2015). Integrins are receptors which are known to promote survival of breast cancer cells and to control metastasis in breast cancer (Felding-Habermann et al., 2001; Parvani et al., 2015). Hedrick et al. also suggested that integrin-mediated signaling pathway as a
novel target for treating high-risk breast cancer patients (Hedrick et al., 2016).

We investigated the possibility that leptin, integrins, or their related genes can be used as prognostic biomarkers. Figure 9(a) and (b) shows that the Kaplan–Meier survival curves of STAT3 and CCND1 which are closely related to leptin, show prognostic significance ($P$-value < 0.05) in cluster0 but not in cluster1. On the other hand, the curves of ITGB1, ITGB3 and PTPN11 show prognostic significance in only cluster1 in Figure 9(c) and (d). These results show that STAT3, CCND1, ITGB1, ITGB3 and PTPN11 can be cluster-specific prognostic biomarkers, which means STAT3/CCND1 and ITGB1/ITGB3/PTPN11 may be mutually exclusive.

Beside leptin and integrin related terms, the relationships between identified GO terms and breast cancer outcome have been investigated. C-type lectin receptor pathway was enriched for biomarkers of cluster 0. C-type lectin receptor is known to regulate adaptive immunity and immunopathogenesis, and has been reported as a new target for cancer immunotherapy (Yan et al., 2015). Ras signaling pathway is known to hyper-activate breast cancer, and it has been studied extensively as a metastasis suppressor (Niemitz, 2013). Elevated Ras signaling was enriched mainly in basal-like and Human Epidermal growth factor Receptor 2-positive (HER2+) subtype tumors, but Ras combined with PIK3CA may be associated with luminal B-like tumor that is ER+ (Wright et al., 2015). PIK3CA was identified in both clusters in Figure 7(a). Hypoxia enriched in cluster0 is associated with risk of metastasis and patient mortality. Since the hypoxic tumor microenvironment affects cancer progression, recent preclinical studies have suggested a therapy with drugs that inhibit hypoxia-inducible factors for good outcomes of patients (Semenza, 2016).

GO terms enriched from biomarkers of cluster1 are somewhat different from those of cluster0. Wnt signaling pathway enriched from genes of cluster1 is known to overexpress in TNBC (Dey et al., 2013). Many researchers studied the pathway as a biomarker for patient’s survival and therapeutic target in TNBC (Holland et al., 2013; Jang et al., 2015). The kinetochore is known to play a role in correcting segregation of chromosomes in mitosis, and its aberration causes chromosomal instability. Recent study has reported that a centromere and kinetochore gene can be used as a marker for prognosis and prediction of response to radiotherapy and chemotherapy (Zhang et al., 2016). Cluster 1 has also Interleukin-1 (IL1) mediated pathway. IL1 is a cytokine that involved in cell growth and death. It is reported to promote cancer cell invasion and metastasis in TNBC (Bouchard et al., 2017).

4 Discussion

The proposed approach is capable of the identification of prognostic genes in the form of networks that provide rich information about
molecular functions related to cancer development. More importantly, our approach identifies different prognostic gene networks for heterogeneous patient sample sets, and allowing us to infer different biological processes related to specific groups of samples by analyzing those networks separately. In addition, hub genes of the union of those gene networks were good prognostic biomarkers that showed better prediction performance in terms of AUC than those of existing algorithms.

Existing studies to overcome the cancer heterogeneity problem generally distinguish samples by known cancer subtype like ER+/- or aim to cluster patient groups with good and poor outcome using various omics data. They study subtype specific biological processes and investigate their relationship with cancer outcome. The proposed approach effectively partitions the samples using principal components of gene expression data. Those groups are not necessarily subtype dependent, but we indirectly proved that these reflect breast cancer heterogeneity by showing that resulting cluster-specific gene networks enrich distinguished biological processes or functions, in Figure 8. Moreover, the clustering step helps better outcome prediction, which additionally supports our claim.

In this study, we clustered samples without prior information, such as ER subtypes, because subtype was not strongly correlated with prognosis (Supplementary Table S1a). To cluster with comprehensive information of whole gene expression, we used PCA on whole genes and selected the two best principal components. However, clustering strategies can vary for different datasets. For example, more principal components can be used, or specific groups of genes can be useful to divide samples that show different expression patterns for those genes. In the future, we will apply our approach to various kinds of cancer data, varying clustering strategies.

Funding
This work was supported by Incheon National University (International Cooperative) Research Grant in 2016 [2016-2295]. This research was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [NRF-2016R1A1A103934135].

Conflict of Interest: none declared.

References