Screening of feature genes in distinguishing different types of breast cancer using support vector machine

Qi Wang
Xudong Liu
Department of Emergency Surgery, Affiliated Hospital of Inner Mongolia Medical University, Hohhot, People’s Republic of China

Objective: To screen the feature genes in estrogen receptor-positive (ER+) breast cancer in comparison with estrogen receptor-negative (ER−) breast cancer.

Methods: Nine microarray data of ER+ and ER− breast cancer samples were collected from Gene Expression Omnibus database. After preprocessing, data in five training sets were analyzed using significance analysis of microarrays to screen the differently expressed genes (DEGs). The DEGs were further analyzed via support vector machine (SVM) function in e1071 package of R to construct a SVM classifier, the efficacy of which was verified by four testing sets and its combination with training sets using a leave-one-out cross-validation. Feature genes obtained by SVM classifier were subjected to function- and pathway-enrichment via the Database for Annotation, Visualization and Integrated Discovery and KEGG Orthology Based Annotation System, respectively.

Results: A total of 526 DEGs were screened between ER+ and ER− breast cancer. The SVM classifier demonstrated that these genes could distinguish different subtype samples with high accuracy of larger than 90%, and also showed good sensitivity, specificity, positive/negative predictive value, and area under receiver operating characteristic curve. The inflammatory and hormone biological processes were the common enriched results for two different function analyses, indicating that the inflammatory (ie, IL8) and hormone regulation (ie, CGA) genes may be the involved feature genes to distinguish ER+ and ER− types of breast cancer.

Conclusion: The gene-expression profile data can provide feature genes to distinguish ER+ and ER− samples, and the identified genes can be used for biomarkers for ER+ samples.

Keywords: classification, differentially expressed genes, biomarker

Introduction

Breast cancer is the most common invasive cancer in females worldwide, with an estimated 232,670 newly diagnosed cases and approximately 40,000 deaths in 2014 in the USA.1 Breast cancer is a hormone-dependent malignancy. At their primary diagnosis, approximately 75%–80% of breast cancer patients present as estrogen receptor-positive (ER+), while 20%–30% are estrogen-negative (ER−).2 It is reported that ER− breast cancer is related with poor prognosis, whereas breast cancer patients who are ER+ have a favorable outcome.3 This indicates the importance to distinguish between these two different subtypes of breast cancer, with the aim to provide prognosis and guide targeted treatment.

Traditional classification based on the histochemical analysis of ER expression is often limited, and does not have the ability to discern subtle differences in different...
subtypes of breast cancer. Thus, molecular identification is advocated. For example, Lim et al demonstrated that the lysine-specific demethylase 1 is highly expressed in ER− breast cancer, while Mehta et al showed that the forkhead box protein A1 is an independent prognostic marker for ER+ breast cancer. However, the research on the genes that could distinguish the different subtypes of breast cancer is so limited and needs further study.

Recent studies have shown that gene-expression profile generated by high-throughput platforms may provide comprehensive molecular characteristics of the tumors and may be informative for tumor classifications. For example, Parker et al have identified a 50-gene transcriptional signature and demonstrated that they have a good prognosis performance for “intrinsic” subtypes of breast cancer (luminal A, luminal B, HER2-enriched, and basal-like). Haibe-Kains et al reported a three-gene-expression model to classify tumors into four molecular entities (ER+/HER2−/low proliferative, ER+/HER2−/high proliferative, HER2+, and ER−/HER2−), which displays relatively less prognosis ability compared to the 50-gene transcriptional signature. However, a gene model to specifically distinguish ER+ and ER− breast cancer remains poorly investigated.

Several data-mining technologies have recently been developed to accomplish feature gene extraction and selection, among which the support vector machine (SVM) algorithm performs at a higher power in two categories of classification. In the present study, we used the SVM to analyze the biomarkers for two subtypes of breast cancer: ER− and ER+ utilizing gene-expression profiling data.

**Materials and methods**

**Microarray data and data preprocessing**

Gene-expression data of breast cancer were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). The following criteria were used to screen out the appropriate gene-expression data: 1) samples were not treated by any medicines; 2) all samples were breast cancer samples; and 3) samples were classified by ER status. Nine datasets were finally included, consisting of 1,289 samples, in which five expression profiles (626 samples, 425 ER+ and 201 ER−) were randomly assigned to training sets and the other four expression profiles (663 samples, 492 ER+ and 171 ER−) were used for testing sets (Table 1). The raw downloaded data first underwent background correction, log2 transformation, and then quantiles normalization using Affy package in R.

**Screening of differentially expressed genes**

To select the differentially expressed genes (DEGs) in ER+ samples compared with ER− samples after data preprocessing, significance analysis of microarrays package in R (www.r-project.org) was utilized. Genes with false discovery rates (FDRs) estimated by permutation method to be less than 0.05 and log2 fold change (FC) >1 were considered as DEGs.

**Sample classification using SVM classifier**

It was not certain whether the selected DEGs could distinguish the two types of breast cancers well; thus, SVM was used for this determination to build models based on “training” data and search for similar patterns in “testing” data. Based on the normalized expression values of 526 DEGs identified using training data, a SVM classifier was constructed via SVM function in e1071 package of R (www.r-project.org) with the nonlinear radial basis function as the kernel and penalty functions set at 1,000. The predictive results of the SVM model for the training set itself were evaluated by a leave-one-out cross-validation method.

**Table 1** Summary of the nine included microarray data

<table>
<thead>
<tr>
<th>Datasets ID</th>
<th>Number of samples</th>
<th>ER+</th>
<th>ER−</th>
<th>Average age (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEO-3494</td>
<td>241</td>
<td>209</td>
<td>32</td>
<td>62.004</td>
</tr>
<tr>
<td>E-GEO-24185</td>
<td>94</td>
<td>56</td>
<td>38</td>
<td>49.03</td>
</tr>
<tr>
<td>E-GEO-22597</td>
<td>82</td>
<td>37</td>
<td>45</td>
<td>51.23</td>
</tr>
<tr>
<td>E-GEO-22093</td>
<td>82</td>
<td>41</td>
<td>41</td>
<td>48.51</td>
</tr>
<tr>
<td>E-GEO-45255</td>
<td>127</td>
<td>82</td>
<td>45</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>626</td>
<td>425</td>
<td>201</td>
<td>52.69</td>
</tr>
<tr>
<td>Testing sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEO-4922</td>
<td>245</td>
<td>211</td>
<td>34</td>
<td>62.12</td>
</tr>
<tr>
<td>E-GEO-32518</td>
<td>71</td>
<td>40</td>
<td>31</td>
<td>47.69</td>
</tr>
<tr>
<td>E-GEO-23988</td>
<td>61</td>
<td>32</td>
<td>29</td>
<td>48.69</td>
</tr>
<tr>
<td>E-GEO-2034</td>
<td>286</td>
<td>209</td>
<td>77</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>663</td>
<td>492</td>
<td>171</td>
<td>52.83</td>
</tr>
</tbody>
</table>

**Abbreviations:** ER+, estrogen receptor-positive; ER−, estrogen receptor-negative.
where one sample in n samples was randomly selected as the testing set, and the other n–1 samples were regarded as the training set. The error rate (1– accuracy), when every single sample in the training sets has been used in the testing set, is the accuracy reference of the SVM classifier. The lower the error rate is, the more accurate the classifier is. Besides accuracy, another five indices were also utilized: sensitivity (Se), specificity (Sp), positive predictive value, negative predictive value, and area under receiver operating characteristic curve. For Se and Sp, $P=0.5$ was the cutoff criteria; while area under the curve was a comprehensive assessment criteria. Subsequently, the accuracy of the SVM classifier was further verified using testing sets and the combined datasets according to the leave-one-out cross-validation method. If all of the earlier mentioned results suggest that the construed SVM classifier exhibits high reliability, DEGs collected from the training sets will be regarded as the feature genes to distinguish the two subtypes of breast cancer.

**Function- and pathway-enrichment of feature genes**

The feature genes were subjected to enrichment analysis to identify their roles in breast cancer. The Database for Annotation, Visualization and Integrated Discovery was used for function enrichment,20 while KEGG Orthology Based Annotation System was applied for pathway-enrichment using hypergeometric distribution algorithm.21 $P<0.05$ was the threshold for the enriched terms.

**Results**

**Screening for DEGs**

A total of 526 DEGs, consisting of 239 upregulated ones and 287 downregulated ones, were identified in ER+ samples, comparing with ER− samples in the five training datasets.

**Sample classification using SVM classifier**

Using the normalized expression values of DEGs, a SVM classifier was constructed (Figure 1). After that, the accuracy of this classifier was detected. For training, testing, and the combined datasets, two (one ER− and one ER+), 29 (16 ER− and 13 ER+), and 22 (seven ER− and 15 ER+) samples were wrongly classified by the SVM classifier, respectively. However, the accuracies were all larger than 90% (99.7% [99.5% for ER− and 99.8% for ER+], 95.6% [90.6% for ER− and 97.3% for ER+], and 98.2% [98.1% for ER− and 98.4% for ER+], respectively), indicating the reliability of

![Figure 1](https://www.dovepress.com/)

**Figure 1** Classification of three sample datasets by constructed support vector machine classifier.

**Notes:** (A) Six hundred and twenty-six samples for training; (B) 663 samples for testing; (C) 1,289 combined samples for testing. (Ab, Bb, and Cb) indicate the sample distribution for ER+ and ER−. (Ab, Bb, and Cb) indicate the scatterplot of the classification, in which black dots represent ER− while red dots represent ER+ breast cancer samples.

**Abbreviations:** ER+, estrogen receptor-positive; ER−, estrogen receptor-negative.
the classifier. Moreover, the results of Se, Sp, positive predictive value, negative predictive value, and area under receiver operating characteristic curve of the SVM classifier showed that it could not only distinguish training datasets, but also testing datasets well (Table 2; Figure 2).

Significantly related functions of feature genes
A total of eight biological functions were enriched by the feature genes (Table 3), among which the response to inorganic substance was the most significant. Furthermore, the majority of features genes were collected in several biological functions, including response to organic substance, cell–cell signaling, response to wounding, behavior, and inflammatory response, each of which accounting for larger than 10%.

Significantly related pathways of feature genes
Eight pathways of the feature genes were enriched, of which, cytokine–cytokine receptor interaction was the most significant one, involving 17 genes. The other pathways included drug metabolism, GnRH signaling pathway, etc (Table 4).

Discussion
In present study, ER+ and ER− types of breast cancer samples were investigated to screen the DEGs, which were further used for SVM classifier training. The SVM classifier could not only distinguish the training dataset, but could also distinguish the testing and combined datasets well, with accuracy higher than 90%. Thus, the DEGs could be considered as feature genes for ER+ and ER− types of breast cancer. Subsequently, function- and pathway-enrichment analyses were conducted, in which the inflammatory- and hormone-related biological process were the common results for these two different analyses, indicating these inflammatory (ie, interleukin-8 [IL8]) and hormone regulation (ie, glycoprotein hormones, alpha polypeptide, CGA) genes may be important for distinguishing ER+ and ER− types of breast cancer.

It is well reported that cytokine IL8 plays an important role in malignant tumor progression. IL8 is highly expressed in breast cancer and is associated with an accelerated clinical course, a higher tumor load, and the presence of distant metastasis, ultimately leading to poor survival. The depletion of IL8 expression may promote the cell cycle arrest and inhibit migration and invasion in breast cancer cells, causing high response to chemotherapy. Thus, IL8 may be a biomarker for distinguishing subtypes of breast cancer because of the lower survival in ER− type of breast cancer. This hypothesis has been demonstrated by several studies. For example, Lin et al showed that IL8 is lowly expressed in ER+ whereas highly expressed in ER− cells. Specifically, knockdown of IL8 significantly reduces the cell invasion by suppressing the PI3K/Akt/NF-κB/integrin β3 pathway in ER− breast cancer cell lines, and the microvessel density and neutrophil infiltration into the tumors in vivo. Furthermore, exogenous addition of ERα in ER− cells may also achieve the goal of downregulating IL8 expression. These findings all suggest the negative relationship between IL8 and ER status, which was also proved in our study (IL8 was downregulated in ER+ samples, log2 FC =−2.25, FDR =1.04×10−70).

CGA codes for the common alpha subunit of four glycoprotein hormones (chorionic gonadotropin, luteinizing

**Table 2** Effect evaluation of the support vector machine classifier on training and testing datasets

<table>
<thead>
<tr>
<th></th>
<th>Number of samples</th>
<th>Correct rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>626</td>
<td>0.9968</td>
<td>0.9976</td>
<td>0.9950</td>
<td>0.9976</td>
<td>0.9950</td>
<td>0.999</td>
</tr>
<tr>
<td>Testing</td>
<td>663</td>
<td>0.9563</td>
<td>0.9736</td>
<td>0.9064</td>
<td>0.9677</td>
<td>0.9226</td>
<td>0.816</td>
</tr>
<tr>
<td>Combined</td>
<td>1,289</td>
<td>0.9829</td>
<td>0.9836</td>
<td>0.9812</td>
<td>0.9923</td>
<td>0.9605</td>
<td>0.890</td>
</tr>
</tbody>
</table>

Abbreviations: AUROC, area under receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.
### Table 3: Significantly enriched biological functions of feature genes

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>P-value</th>
<th>Enriched genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>gO:0010035, response to inorganic substance</td>
<td>24</td>
<td>1.28×10⁻⁷</td>
<td>SYT1, ERBB4, CRAB, S100A7, EEF1A2, ALDOB, TRPA1, NR4A2, GGH, SOD2, IGFBP2</td>
</tr>
<tr>
<td>gO:0007267, cell–cell signaling</td>
<td>43</td>
<td>1.02×10⁻⁶</td>
<td>LALBA, CGA, SYT1, CXCL5, NDP, S100A9, CXCL9, CCL8, CACNB2, GABBR2, IGFBP4</td>
</tr>
<tr>
<td>gO:0009611, response to wounding</td>
<td>37</td>
<td>1.2×10⁻⁵</td>
<td>CXCL1, TF, S100A8, ERBB2, S100A9, CXCL9, CCL8, CXCL11, CDH3, IGFBP4</td>
</tr>
<tr>
<td>gO:0006954, inflammatory response</td>
<td>27</td>
<td>1.29×10⁻⁵</td>
<td>CXCL1, TF, S100A8, S100A9, CXCL9, CCL8, CXCL11, FOS, IL17B, IGFBP4</td>
</tr>
<tr>
<td>gO:0010038, response to metal ion</td>
<td>16</td>
<td>1.3×10⁻⁵</td>
<td>SYT1, ALDOB, GGH, PCSK1, FGG, CCND1, PLAZA2A, GRIA2, FGB, CYBRD1, IGFBP4</td>
</tr>
<tr>
<td>gO:0007610, behavior</td>
<td>34</td>
<td>1.38×10⁻⁵</td>
<td>CXCL1, ADCY1, CCL8, ADCY9, UCHL1, CCL9, TRH, ZIC1, IGFBP4</td>
</tr>
<tr>
<td>gO:0010033, response to organic substance</td>
<td>45</td>
<td>1.94×10⁻⁵</td>
<td>CGA, TF, KYN, ADCY1, CYP1B1, ERBB4, IL6ST, ERBB2, ARNT2, ALDOB, IGFBP2</td>
</tr>
<tr>
<td>gO:0010817, regulation of hormone levels</td>
<td>17</td>
<td>2.2×10⁻⁵</td>
<td>KLK6, CGA, CYP1B1, TBX3, FOXA1, AFP, PCSK1, DHR52, WNT4, SERPINA6, SNAP25</td>
</tr>
</tbody>
</table>

**Abbreviation:** GO, gene oncology.

Figure 3: Significantly enriched functions of feature genes.

**Note:** Different colors represent different biological functions and the percentages are the ratios of genes in a function.

### Table 4: Significantly enriched pathways of feature genes

<table>
<thead>
<tr>
<th>ID</th>
<th>Pathway</th>
<th>P-value</th>
<th>Enriched genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04060</td>
<td>Cytokine–cytokine receptor interaction</td>
<td>2.03×10⁻²</td>
<td>CXCL1, IL1R2, IL2RA, IL8, CXCL5, IL6ST, CXCL9, CCL8, CXCL11, CCL18, CXCL14, CCL20, IL20RA, BMR1B, LTBP3, GRIA2, FGB, CYBRD1, IGFBP4</td>
</tr>
<tr>
<td>hsa00982</td>
<td>Drug metabolism</td>
<td>2.05×10⁻²</td>
<td>GSTA1, FMOS, FMOS, UGT2B4, CYP2A6, CYP2A7, GSTP1</td>
</tr>
<tr>
<td>hsa04912</td>
<td>GnRH signaling pathway</td>
<td>2.07×10⁻²</td>
<td>CGA, ADCY1, PLCG2A, ADCY9, MAPK14, CAMK2B, CALML5, CACNA1D, ITPR1</td>
</tr>
<tr>
<td>hsa00232</td>
<td>Caffeine metabolism</td>
<td>2.28×10⁻²</td>
<td>NAT1, CYP2A6, CYP2A7</td>
</tr>
<tr>
<td>hsa04062</td>
<td>Chemokine signaling pathway</td>
<td>3.00×10⁻²</td>
<td>CXCL1, ADCY1, VAV3, IL8, CXCL5, CXCL9, CCL8, CXCL11, CCL18, CXCL14, CCL20, IL20RA, BMR1B, LTBP3, GRIA2, FGB, CYBRD1, IGFBP4</td>
</tr>
<tr>
<td>hsa04914</td>
<td>Progesterone-mediated oocyte maturation</td>
<td>3.01×10⁻²</td>
<td>PGR, IGFR1, ADCY1, ABCY9, MAPK14, BUB1, IGF2, CDC25A</td>
</tr>
<tr>
<td>hsa04110</td>
<td>Cell cycle</td>
<td>3.03×10⁻²</td>
<td>CCNE1, CCND1, CDC45, CDKN2A, BUB1, TTK, CDC20, SFN, MCM4, CDC25A</td>
</tr>
<tr>
<td>hsa00380</td>
<td>Tryptophan metabolism</td>
<td>4.98×10⁻²</td>
<td>WARS, KYN, CYP1B1, IDO1, TPH1</td>
</tr>
</tbody>
</table>
hormone, follicle stimulating hormone, and thyroid stimulating hormone) that have a cystine knot motif formed by three of the five disulfide bonds.\textsuperscript{32,33} This cystine knot motif is known to be a characteristic feature of growth factor and, thus, the expression of CGA may be related with the development of cancer, which was also demonstrated in previous studies.\textsuperscript{34,35} However, compared with the beta subunit of glycoprotein hormones,\textsuperscript{36} the alpha subunit may be a marker of tumors with low aggressiveness, eg, ER+ but not ER– cancer cells.\textsuperscript{37,38} In this study, we also found the upregulated expression of CGA in ER+ breast cancer patients (log2 FC =2.63, FDR =2.82×10\textsuperscript{−14}).

Despite the ideal classification of different types of breast cancer samples and satisfactory accuracy, the SVM classifier showed decreased recognizing ability. The potential reasons for this are: 1) breast cancer is one of the tumors which vary between individuals, and this variation among samples will affect gene expressions; and 2) samples used for training and testing were obtained from different experiments, which allows for some personal error. This kind of error can hardly be eliminated by normalization of the data. However, all the other indices showed reliability of our classifier by SVM method.

**Conclusion**

Based on a set of gene-expression profiles, 526 DEGs were identified in ER+ samples in comparison with ER– samples, which were further used for SVM classifier construction. After being tested using the other microarray, the SVM classifier showed satisfactory efficacy. The selected feature genes (such as IL8 and CGA) could well distinguish those two subtypes of breast cancer. However, further experimental studies are needed to confirm the values of other involved genes.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


