A Coupled Three-Step Network-Based Approach to Identify Genes Associated with Breast Cancer

Michael Netzer*, Xiaocong Fang†, Michael Handler*, Christian Baumgartner*

*Research Group for Clinical Bioinformatics, Institute of Electrical, Electronic and Bioengineering, UMIT, 6060 Hall in Tirol, Austria
†Zhongshan Hospital, Fudan University
Email: michael.netzer|michael.handler|christian.baumgartner@umit.at
20032 Xuhui District, Shanghai, China
fangxiaocong@gmail.com

Abstract—New biomarker candidates for breast cancer recurrence are urgently needed to provide patients an optimal treatment and avoid “overtreatment” where patients needlessly suffer from the toxic side effects of chemotherapy. In this work, we present a new network-based approach to identify biomarker candidates for breast cancer recurrence. Our coupled three-step strategy first selects relevant genes by using a filter approach. In the second step, we infer a new type of network where the number of edges of each vertex (i.e., degree) represents the predictive value of the underlying gene. In the third step, we conduct a database search for biomedical interpretation of our findings. Using a breast cancer microarray dataset, we could verify top-ranked genes associated with breast cancer and other pathophysiological processes.

Index Terms—biomarker discovery; microarray; feature selection; networks; breast cancer

I. INTRODUCTION

Recent biotechnological advances in the “omics” sciences such as microarrays have led to high amounts of data. In particular, this data is characterized by a high number of features (e.g., genes) and a small number of samples (large p, small n problem) [1]. To identify highly predictive biomarker candidates, sophisticated feature selection approaches, including filters and wrappers, are required [2]. Wrapper approaches [3] use learning algorithms (i.e., a classifier if the dependent variable is categorical) and a search strategy to identify sets of highly discriminating features. Filter approaches that calculate a measure for every feature representing its predictive ability are independent of a classifier and generally have less extensive computational costs [4].

Advances of computer systems have also led to extensive biological network analysis [5]–[7], increasing understanding of interactions between genes, proteins, and metabolites.

In this work, we introduce a new three-step network-based approach to identify genes related to breast cancer recurrence. Our coupled three-step approach first selects relevant genes by using a filter approach. In the second step, we infer a new type of network where the number of edges of each vertex (i.e., degree) represents the predictive value of the underlying gene. In the third step, we conduct a database search for biomedical interpretation of our findings. Using a breast cancer microarray dataset, we could verify top-ranked genes associated with breast cancer and other pathophysiological processes.

II. MATERIALS AND METHODS

A. Data

We used 59 preprocessed microarray spectra (platform GPL1390) from the study of Hoadley et al. [9] available at the GEO database [10] (n_s = 59). The data includes 29 controls and 30 cases (breast cancer recurrence within 36 months). In our experiments, we only use genes where the ensembl gene id is available (n_f = 11638).

B. Computational approach

Given is a set of tuples (dataset) T = {x_i, y_i}^{n_s}_{i=1} with x_i ∈ R^{n_f} and y_i ∈ {case, control}, where n_f is the number of features, n_s is the number of samples, and y is the set of class labels. Recently, we developed a new supervised approach to infer networks based on the ratios between metabolites [11]. Therefore, we first calculated all ratios R between features F, where r_{ij} = |log_2 (\frac{F_i}{F_j})| with i > j, and f ∈ F, r ∈ R.

Note that the logarithm induces symmetry of the ratios and their reciprocals, respectively. We then created a graph G with:

\[ G_{ij} = \begin{cases} 1 & \text{if } |s_{ij}| > \tau \\ 0 & \text{else} \end{cases} \]

(1)

for i, j ∈ 1, …, n_f, where \( \tau \) is a defined threshold. The score s representing the discriminatory ability was calculated using the BI filtering method [12]. Briefly, BI combines a discrimination measure (DA*), the coefficient of variation (CV) and \( \Delta_{\text{change}} \) representing the strength of the change. For the unpaired BI, we define \( \Delta = \frac{\bar{\pi}_{ref}}{\bar{\pi}_{case}} \), where \( \bar{\pi} \) is the mean of the comparison group (e.g., case) and \( \bar{\pi}_{ref} \) is the mean of the reference group (e.g., control).

In our recent work, we could show that this new type of network, based on the ratios, outperforms other networks as

Consequently, many patients are over-treated needlessly, suffering from toxic side effects of chemotherapy [8].

Section “Materials and Methods” describes our new three-step network-based approach. In the “Results and Discussion” section, we show our results for an example dataset. Finally, we conclude and discuss our method and findings.
correlation networks in terms of accuracy (see also Netzer et al., 2011). However, this approach has several limitations: i) Creating the network for \( n \) features results in \( \frac{n(n-1)}{2} \) comparisons and BI calculations (\( \mathcal{O}(n^2) \)); ii) by definition the values for \( f \) must be positive to calculate the logarithms and \( \tau_{ref} \) must be \( \neq 0 \) to calculate the BI scores.

In particular, when dealing with standard normal distributed data (\( N(0,1) \)) we have a high number of features \( (n > 10,000) \) including negative values for \( f \) and \( \tau_{ref} \) is close to zero.

Therefore, in this work, we propose a new generic three-step strategy to overcome the aforementioned restrictions for standard normal distributed datasets with a high number of features (e.g., preprocessed microarray data):

**Step 1:** In order to reduce the number of features we first perform a feature selection. Therefore, we use a filter method and calculate the score \( s \) representing the discriminatory ability of each feature. We remove all features \( f \) with a score less than a defined threshold \( (s_f < \tau) \). Finally, we obtain our reduced dataset \( T_r \).

**Step 2:** Given a standard normal distributed dataset we calculate all differences \( D \) between features \( F \) in \( T_r \), where \( d_{ij} = |f_i - f_j| \) with \( i > j \), and \( f \in F, d \in D \). Similar to equation 1 we finally construct the graph with

\[
G_{ij} = \begin{cases} 
1 & \text{if } s_{ij} > \tau \\
0 & \text{else}, 
\end{cases} \quad (2)
\]

In our experiments, we used the information gain [13] to calculate the score \( s \) on the distances \( d \) for step 1 and 2. The information gain \( IG \) of a feature \( f \) is given by [14]

\[
IG = H(Y) - H(Y|X), \quad \text{(3)}
\]

\[
H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)), \quad \text{(4)}
\]

\[
H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x) p(y|x) \log_2(p(y|x)), \quad \text{(5)}
\]

where \( H(Y) \) denotes the entropy for \( Y \) (class variable) and \( H(Y|X) \) is the entropy of \( Y \) after observing \( X \).

The information gain easily allows to identify features with no or less discriminatory ability. Therefore, we set \( \tau = 0 \) (filtering threshold) to remove features with no information regarding the class attribute. In addition, we also used the information gain because it can also deal with values \( \leq 0 \).

**Step 3:** After the network is inferred the genes are ranked according to the topological descriptor of the vertices. In this study we ranked the genes according to their degree (i.e., number of edges).

To verify and interpret our findings a database search to multiple repositories such as the Database for Annotation, Visualization and Integrated Discovery (DAVID) [15] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database was conducted [16].

The feature selection was performed using Weka [17]. We used R [18] to implement the network approach methods. There exist several R packages for handling and analyzing graphs such as graph [19], igraph [20], QuACN [21] and BioNet [22]. The overall workflow including data import, feature selection, network inference and pathway analysis is depicted in Fig. 1.

The used hardware platform was an Intel Centrino 2x-1.83 GHz PC with 2048 MB RAM.

### III. Results and Discussion

This work aims at identifying new genes related to breast cancer recurrence. The distribution of the used dataset is shown in Fig. 2. After feature selection of step 1 a total of 156 genes yielded an information gain > 0. The resulting network after applying step 2 to the reduced dataset is shown in Fig. 3. The ten top ranked genes using the degree (i.e., number of edges) are listed in Table I.

The related pathophysiological processes of the identified gene set are shown in Table II.

Kallkreins gene family are a group of serine proteases and known to be involved in several endocrine related malignancies.
Kallikrein gene 7 (KLK7) was first reported in 1991 in the desquamation of stratum corneum [24] and later reported to be highly upregulated in ovarian carcinomas [25]. Recent studies have reported that KLK7 can be upregulated primarily by estrogens and glucocorticoids [26], [27]. The expression of the KLK7 gene was supposed to be the most independent prognostic marker for the survival of patients with breast cancer [27].

Salivary amylase alpha 1 (AMY1) is involved in the starch and sucrose metabolism. It was reported to be highly expressed in individuals with high starch diets [28] and highly activated under psychosocial stress [29]. There is no existing evidence which supports the association between AMY1 and breast cancer, however, recent study reported AMY1 is an important modulator of cAMP-dependent protein kinase (PKA) which has versatile functions in cells [30].

The PH domain of PHLDA2 can compete with the PH domain of some other proteins, thereby interfering with their binding to phosphatidylinositol 4,5-bisphosphate (PIP2) and phosphatidylinositol 3,4,5-triphosphate (PIP3) in membrane lipids thus involved in various biological processes [15].

SnrpC belongs to the U1 small nuclear ribonucleoprotein C family and was reported to be involved in the splicing of mRNA [31]. The peptidylarginine deiminases 2 (PAD2) is a member of PADs which catalyze citrullination by converting arginine residues of proteins [32], [33]. It was also reported to play a role in inflammatory response, cell apoptosis [34], [35] as well as the gene regulation in the mammary glands [36].

Wee1 is another gene which was previously reported to be involved in tumourigenesis [37], [38] as well as the signaling regulation in breast cancer stem cells [39]. It was suggested to act as a tumor suppressor via regulating Cyclin and cyclin-dependent kinase complexes [37].

SOCS family proteins are part of a classical negative feedback system that regulates cytokine signal transduction of which SOCS2 appears to regulate the growth hormone/IGF1 signaling pathway [15], [40]. It is also suggested to play a role in the oncogenesis of ovary and mammary gland [41].

SOX5 is one of the high-mobility group (HMG) which has been recognized as a key player in the regulation of embryonic development [42] and in the determination of cell fate [43] reported to be involved in the progression of glioma and prostate cancer [44], [45].

KIAA0494 was supposed to play a role in the splicing of eukaryotic pre-mRNAs [46] but there are no more studies provided on its effect on tumorigenesis. CTP synthase (CTPS) plays a predominant role in CTP synthesis by converting UTP to CTP thus controlling cell proliferation, differentiation and apoptosis [47]. Previous studies have demonstrated that CTPS depletion resulted in stabilization of wild-type p53 and showed antitumor effects in breast cancer cells [48].

Among the top 10 most related genes, we found that half of them had been previously reported to play roles in tumorigenesis, such as KLK7, Wee1, SOCS, SOX5 and CTPS. Furthermore, some had been studied in breast cancer (KL7, Wee1, SOCS, CTPS). However, the others were first suspected to be involved in the development or progression of breast cancer. Our results revealed that it is possible that these genes provide new targets for the control of breast cancer. However, further studies are warranted and essential to verify and validate these promising findings.

IV. Conclusions

In this work, we introduced a new network-based approach to identify biomarker candidates for breast cancer recurrence. Our main contribution is to propose a new workflow that
TABLE II
THE RELATED PATHOPHYSIOLOGICAL PROCESSES OF THE TOP 10 RANKED GENES. THE ASSOCIATED GENE NAMES CAN BE FOUND IN TABLE I.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Related pathophysiological processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Play a role in the desquamation of the skin, the stratum corneum. Up-regulated in by estrogens and glucocorticoids.</td>
</tr>
<tr>
<td>2</td>
<td>involved in starch and sucrose metabolism, dafily secretion, carbohydrate digestion and absorption.</td>
</tr>
<tr>
<td>3</td>
<td>compete to bind phosphoinositides in the membrane lipids with a broad specificity in various biological processes</td>
</tr>
<tr>
<td>4</td>
<td>associated with snRNP U1 and involved in the spliceosome pathway.</td>
</tr>
<tr>
<td>5</td>
<td>catalyzes the deamination of arginine residues of proteins</td>
</tr>
<tr>
<td>6</td>
<td>participate in the cell cycle with an increased synthesis during S and G2 phases</td>
</tr>
<tr>
<td>7</td>
<td>a negative regulator in the growth hormone/IGF1 signaling pathway</td>
</tr>
<tr>
<td>8</td>
<td>binds specifically to the DNA sequence 5’-AACAAT-3’, overexpressed in glioma and prostate tumor</td>
</tr>
<tr>
<td>9</td>
<td>involved in the splicing of eukaryotic pre-mRNAs</td>
</tr>
<tr>
<td>10</td>
<td>catalyzes the ATP-dependent amination of UTP to CTP, play a role in pyrimidine metabolism</td>
</tr>
</tbody>
</table>

includes the filtering of relevant genes and inferring the network based on an information entropy measure. Our three-step strategy selected in the first step relevant features using the information gain. In the second step we inferred a new type of a network where the number of edges of each vertex (i.e., degree) represents the predictive ability of the underlying feature. Finally we used the DAVID and KEGG databases to verify and interpret top ranked genes.

Using our breast cancer microarray dataset from the GEO database we could identify a set of known and unexpected genes associated with breast cancer and other pathophysiological processes. The proposed generic method can also be applied to other biomedical questions (e.g., other diseases) or types of data such as metabolomic datasets.

ACKNOWLEDGMENT

This work was supported by the Austrian Genome Research Program GEN-AU (Bioinformatics Integration Network, BIN III) and partly by the Tiroler Wissenschaftsfond.

REFERENCES


