IDENTIFYING RELIABLE SUBNETWORK MARKERS IN PROTEIN-PROTEIN INTERACTION NETWORK FOR CLASSIFICATION OF BREAST CANCER METASTASIS

Junjie Su and Byung-Jun Yoon

Dept. of Electrical Engineering and Computer Engineering, Texas A&M University
College Station, TX 77843, USA

ABSTRACT

Due to the inherent measurement noise in microarray experiments, heterogeneity across samples, and limited sample size, it is often hard to find reliable gene markers for classification. For this reason, several studies proposed to analyze the expression data at the level of groups of functionally related genes such as pathways. One practical problem of these pathway-based approaches is the limited coverage of genes by known pathways. To overcome this problem, we propose a new method for identifying effective subnetwork markers by overlaying the gene expression data with a genome-scale protein-protein interaction network. Experimental results on two independent breast cancer datasets show that the subnetwork markers lead to more accurate classification of breast cancer metastasis and are more reproducible than both gene and pathway markers.

Index Terms— Protein-protein interaction (PPI) network, subnetwork markers, cancer classification.

1. INTRODUCTION

Given the high throughput genomic data from microarray studies, one challenge is to find bio-markers associated with disease states. Significant amount of work has been done to identify gene markers that can be used for disease classification. However, due to the inherent measurement noise in microarray experiments, heterogeneity across samples, and limited sample size, it is often hard to find reliable gene markers. Moreover, the gene markers are often selected independently although some of them may be actually functionally related. For this reason, the selected gene markers often contain redundant information that may lead to degradation of classification performance.

To address this problem, several studies proposed to interpret the expression data at the level of groups of genes that are functionally related, for example, known biological pathways [1, 2, 3, 4]. One advantage of using pathway markers is that we are able to reduce the effect of the measurement noise and that of the correlations between genes within the same pathway. This can be achieved by capturing the overall expression changes of a given pathway by jointly analyzing the expression levels of its member genes. It has been demonstrated that pathway markers are more reliable compared to gene markers and lead to better or comparable classification performance. Furthermore, pathway markers can provide biological insights into the underlying mechanisms of different disease phenotypes. One practical problem of this pathway-based approach is that the currently known pathways cover only a limited number of genes. This may exclude key genes with significantly expression changes across different phenotypes. Moreover, many pathways in public databases often overlap with each other, which may also introduce correlations between some of the selected pathway markers.

The availability of large protein-protein interaction(PPI) networks provides us a possible way to alleviate these problems. Recently, there have been research efforts to identify subnetwork markers by overlaying the gene expression data with a protein-protein interaction network [5]. In [5], they first found significantly differentially expressed seed genes and then greedily grew the subnetworks from the seeds such that the mutual information between the subnetwork activity scores and the class labels was maximized. It has been shown that subnetwork markers have good reproducibility and lead to more accurate classification.

In this paper, we introduce a new method for identifying effective subnetwork markers from a PPI network by performing a global search for differentially expressed linear pathways using dynamic programming. Overlapping pathways are optimally combined into subnetworks, which are used as bio-markers for classifying breast cancer metastasis. This paper is organized as follows. We first describe the algorithm in the following section. In Sec. 3, we used this algorithm to identify subnetwork markers and evaluate their effectiveness based on two independent breast cancer datasets. We discuss the obtained results and conclude this paper in Sec. 4.

2. IDENTIFICATION OF SUBNETWORK MARKERS

Given a large PPI network, we want to find subnetwork markers whose activity scores are indicative of the disease states of interest. For this purpose, we first need a method for inferring the subnetwork activities and evaluating their discriminative power. Currently, there exist different ways for computing the activity score of a given group of genes [4]. We have shown that the probabilistic inference scheme proposed in [4] outperforms many other existing methods. Thus, we adopt this activity inference scheme for finding subnetwork markers whose activity scores are highly discriminative of the disease states. However, finding the subnetwork markers with maximum discriminative power in a PPI network based on the selected inference method is computationally infeasible. For this reason, we propose an algorithm for identifying effective subnetwork markers which is motivated by a simpler scheme proposed in [6]. This scheme has been shown to be effective in approximately evaluating the discriminative power of pathway markers [4].

The general outline of the proposed algorithm is as follows. Based on the scoring scheme suggested in [6], we first search for differentially expressed pathways using dynamic programming. Then, the top pathways that overlap with each other are optimally combined into a subnetwork based on the activity inference method proposed in [4]. The identified subnetwork is removed from the PPI network, and the above process is repeated to find multiple non-overlapping subnetwork markers. The overall scheme is illustrated in Fig. 1.

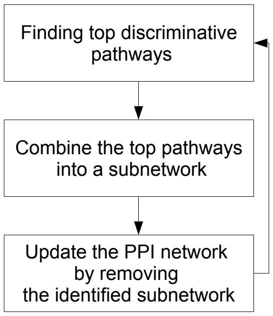
Finding top discriminative pathways
Combine the top pathways into a subnetwork
Update the PPI network by removing the identified subnetwork

Fig. 1. Illustration of the proposed algorithm.

2.1. Evaluating the discriminative power of pathways

A linear pathway $\lambda = \{g_1, g_2, \cdots, g_n\}$ in a given PPI network $\mathcal{G}$ is defined as a group of genes, where g_i and g_{i+1} are connected for $i = 1, \cdots, n-1$. Assume that the expression level x_i of a gene g_i follows the distribution $f_k(x_i)$ under phenotype k, where $k = 1, 2$. The log-likelihood ratio [4] between two phenotypes is computed as follows

$$\alpha_i(x_i) = \log(f_1(x_i)/f_2(x_i)). \tag{1}$$

We evaluate the discriminative power of the gene g_i by computing the t-test statistic score of the log-likelihood ratio $\alpha_i(x_i)$, denoted as $t_\alpha(g_i)$. The discriminative power of the pathway λ can be assessed by taking the average absolute t-score of the log-likelihood ratios of its member genes.

$$S(\lambda) = \sum_{g_i \in \lambda} |t_\alpha(g_i)|/n. \tag{2}$$

The above scoring scheme can be used for finding the top linear pathways in the network $\mathcal{G}$ as we describe in the following section.

2.2. Searching discriminative pathways

Let $\mathcal{G} = (E, V)$ denote the PPI network, where V is the set of nodes (i.e., proteins), E is the set of edges (i.e., protein interactions). Suppose there are N proteins in $\mathcal{G}$, we can represent E as an N-dimensional binary matrix. For any protein pair (v_a, v_b), where $v_a, v_b \in V$, we let $E[v_a, v_b] = 1$, if v_a, v_b are connected; $E[v_a, v_b] = 0$, otherwise. Based on the pathway scoring scheme defined in Sec.2.1, we search for top discriminative pathways using dynamic programming. We define $\lambda(v_i, l)$ as the optimal pathway among all pathways with length l and ending at v_i, whose score is denoted as $s(v_i, l)$

$$s(v_i, l) = \sum_{v_\theta \in \lambda(v_i, l)} |t_\alpha(v_\theta)|.$$

Here, only pathways with length $l \leq L$ are considered. The algorithm is defined as follows.

```
(i)Initialization: $l = 1, \forall v_i \in V$,
        $s(v_i, l) = |t_\alpha(v_i)|$.
(ii) Iteration:
    for $l = 2$ to $L$,
      for $\forall v_i \in V$,
        $s(v_i, l) = \max_{v_j}\{s(v_j, l-1) + t_\alpha(v_i) + \log(E[v_i, v_j])\}$,    (3)
        $v_j^* = \arg\max_{v_j}\{s(v_j, l-1) + t_\alpha(v_i) + \log(E[v_i, v_j])\}$,    (4)
        if $s(v_i, l) > 0$, then
          $\lambda(v_i, l) = \lambda(v_j^*, l-1) \cup \{v_i\}$.
      end
    end
(iii) Termination:
    for $\forall v_i \in V, 1 \leq l \leq L$,
        $S(\lambda(v_i, l)) = s(v_i, l)/l$.    (5)
```

Although the above algorithm finds only the top pathway for every (v_i, l), we can easily modify it to get the top M discriminative pathways. Then, the complexity of this algorithm is $O(ML \cdot N^2)$.

2.3. Combining pathways into subnetwork

Based on (5), we choose the m top scoring pathways $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_m\}$ whose lengths are within a given range $[L_{\min}, L_{\max}]$. Next, the pathways in Λ are combined into a subnetwork $\mathcal{G}_s$ so that its discriminative power $\mathcal{R}(\mathcal{G}_s)$ is locally optimized. This process is carried out as follows.
(i) $\mathcal{G}_s \leftarrow \lambda_1, \mathcal{G}_{temp} \leftarrow \mathcal{G}_s, i = 1$.
(ii) $i = i + 1$; If $\lambda_i \cap \mathcal{G}_s \neq \emptyset, \mathcal{G}_{temp} \leftarrow \mathcal{G}_{temp} \cup \lambda_i$.
(iii) If $R(\mathcal{G}_s) < R(\mathcal{G}_{temp}), \mathcal{G}_s \leftarrow \mathcal{G}_{temp}$; else $\mathcal{G}_{temp} \leftarrow \mathcal{G}_s$.
(iv) Go to (ii) if $i < m$; otherwise, terminate.
In order to estimate the discriminative power of a subnetwork, we used the activity inference method in [4] to infer the actual activity score of $\mathcal{G}_s$. Then, $\mathcal{R}(\mathcal{G}_s)$ is computed as the t-test statistics of the subnetwork activity score.

After obtaining a subnetwork $\mathcal{G}_s$, we removed it from the network $\mathcal{G}$ by setting $E[v_s, v_i] = E[v_i, v_s] = 0, \forall v_s \in \mathcal{G}_s, v_i \in \mathcal{G}$. Then, the whole process was repeated based on the updated network to find additional subnetwork markers.

3. EXPERIMENTAL RESULTS

We demonstrated the effectiveness of the identified subnetworks in two different ways based on two independent breast cancer datasets. First, we evaluated the discriminative power of the subnetwork markers. Second, subnetwork markers were applied to the classification of breast cancer metastasis.

3.1. Datasets

We obtained two independent breast cancer datasets from the large-scale expression studies in [7] (referred as the USA dataset) and [8] (referred as the Netherlands dataset). The USA dataset contains 286 samples with 107 metastasis and 179 metastasis-free samples. The Netherlands dataset contains 295 samples with 79 metastasis and 216 metastasis-free samples. The PPI network has been obtained from [5], which contains 57,235 interactions among 11,203 proteins. Since not all proteins have corresponding genes in the microarray platforms used by the two breast cancer studies, we used the induced network which contains 9,079 proteins and 48,734 interactions for the USA dataset, and 5,541 proteins and 28,034 interactions for the Netherlands dataset.

3.2. Discriminative power of subnetwork markers

For each dataset, 50 subnetwork markers were identified using the proposed method. The parameters were set to $M = 20$, $L_{\max} = L = 8$, $L_{\min} = 5$, and $m = 100$. For the USA dataset, the mean size of the identified subnetwork markers was 19.2 with a standard deviation of 10.2. For the Netherlands dataset, the average was 18 and the standard deviation was 9.6.

To assess the discriminative power of a given subnetwork marker, we computed its activity score using the inference method proposed in [4] and estimated its t-test statistic score. Pathways were sorted in the decreasing order of t-score. Average t-scores of the top $K = 10, 20, 30, 40, 50$ subnetworks are shown in Fig. 2A and Fig. 2B. For comparison, the top 50 pathways in the C2 curated gene sets in MsigDB[1](Molecular Signatures Database), which contains 639 known biological pathways, were identified based on each dataset. Then, the discriminative power of the selected pathway markers was computed in the same way. We also estimated the discriminative power of the top 50 genes covered by the identified subnetwork markers. As we can see from Fig. 2A and Fig. 2B, subnetwork markers are more discriminative than both pathway and gene markers.

To compare the reproducibility of different markers, we identified the top 50 markers based on one dataset and evaluated on the other dataset. Figures 2C and 2D show that the subnetwork markers retain higher discriminative power across different datasets.

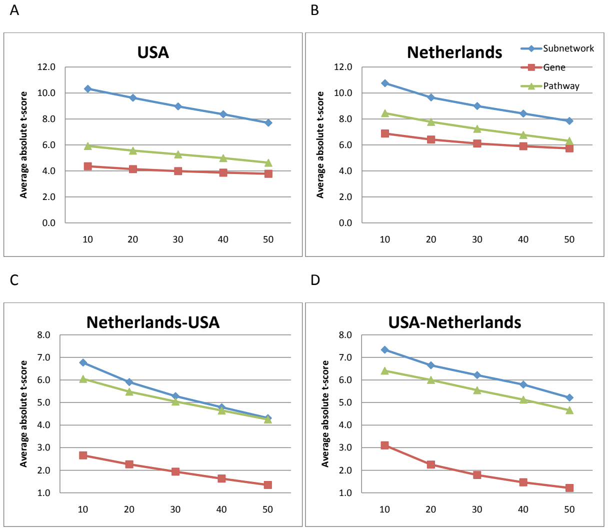

Fig. 2. Discriminative power of different markers. A, B: Markers were identified using a particular dataset and tested on the same dataset. C, D: Markers were identified using one dataset and evaluated based on the other dataset.

3.3. Classification performance

To evaluate the classification performance of the classifiers based on subnetwork markers, we performed the following within-dataset and cross-dataset cross-validation experiments.

In within-dataset experiments, the top 50 subnetwork markers identified using one of the two breast cancer datasets were used to build the classifier. The dataset was divided into five folds of equal size, one of them was withheld as test set and the remaining four were used for training the

[1] http://www.broadinstitute.org/gsea/msigdb/index.jsp

classifier. In the training set, three folds were used to rank the subnetwork markers in the increasing order of p-value and build the classifier based on logistic regression, while one fold was used for feature selection. We started with the top ranked subnetwork marker and built the classifier by adding features sequentially. If the AUC metric [9] on the feature selection set increased, the additional feature was kept; otherwise, we discarded it and tested the remaining ones. The above experiment has been repeated 500 times based on 100 random five-fold splits. The average AUC was reported as the classification performance. Similar experiments have been performed using pathway and gene markers, respectively. The two bar charts on the left of Fig. 3 show that the subnetwork markers significantly improved the classification performance.

To evaluate the reproducibility of subnetwork markers, we performed the following cross-dataset experiments. We identified the top 50 subnetwork markers based on one dataset and performed the same cross-validation experiment which had been used in the within-dataset experiments on the other dataset. The experimental results are shown in the two bar charts on the right of Fig. 3. As before, the classification performance of subnetwork-based classifiers significantly outperform that of the classifiers based on gene or pathway markers.

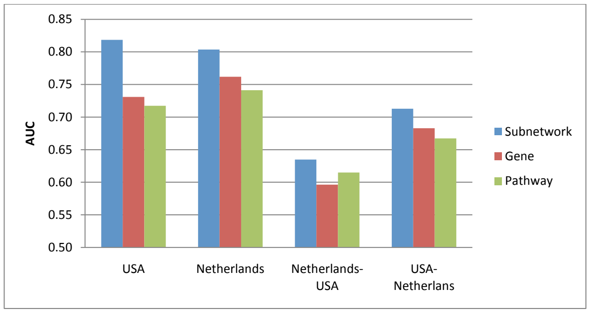

Fig. 3. Classification Performance. Bar charts labeled USA and Netherlands are the results of the within-dataset experiments. Bar charts labeled NethersIands-USA and USA-Netherlands are results of the cross-dataset experiments where markers were identified based on the first dataset and tested based on the second one.

4. DISCUSSION AND CONCLUSION

In this paper, we proposed a new method for identifying effective subnetwork markers in a PPI network. The proposed method finds top scoring pathways using dynamic programming and combines them into a subnetwork to optimize the discriminative power of the resulting subnetwork markers. In this work,the activities of the identified subnetwork markers were inferred using the probabilistic inference scheme proposed in [4]. There are several advantages of the proposed method. First of all, the genome-scale PPI network used in this paper provides a better coverage of the genes in microarray studies compare to pathways obtained from public databases. Second, we construct the subnetworks based on differentially expressed pathways instead of starting from single genes, therefore the result subnetworks may be more robust. Moreover, the probabilistic inference scheme leads to the identification of better subnetwork markers since it can more reliably assess their discriminative powers. As shown in this paper, classifiers based on the subnetwork markers identified using the proposed method achieve higher classification accuracy in both within and cross dataset experiments.

5. ACKNOWLEDGEMENT

We would like to thank the authors of [5] especially H.Y. Chuang for sharing the PPI network.

6. REFERENCES

[1] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, Q. Wang, and S. Rao, “Towards precise classification of cancers based on robust gene functional expression profiles,” *BMC Bioinformatics*, vol. 6, pp. 58, 2005.

[2] J. Tomfohr, J. Lu, and T. B. Kepler, “Pathway level analysis of gene expression using singular value decomposition,” *BMC Bioinformatics*, vol. 6, pp. 225, 2005.

[3] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee, “Inferring pathway activity toward precise disease classification,” *PLoS Comput. Biol.*, vol. 4, pp. e1000217, Nov 2008.

[4] J. Su, B. J. Yoon, and E. R. Dougherty, “Accurate and reliable disease classication based on probabilistic inference of pathway activity ,” Submitted.

[5] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, “Network-based classification of breast cancer metastasis,” *Mol. Syst. Biol.*, vol. 3, pp. 140, 2007.

[6] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, “Discovering statistically significant pathways in expression profiling studies,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 13544–13549, Sep 2005.

[7] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,” *Lancet*, vol. 365, pp. 671–679, 2005.

[8] L. J. van ’t Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, pp. 530–536, Jan 2002.

[9] T. Fawcett, “An introduction to ROC analysis. Patt Recog Letters,” *Patt Recog Letters*, vol. 27, pp. 861–874, Jun 2006.