Su et al. BMC Bioinformatics 2010, 11(Suppl 6):S8
http://www.biomedcentral.com/1471-2105/11/S6/S8

BMC Bioinformatics

PROCEEDINGS Open Access

Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network

Junjie Su[1], Byung-Jun Yoon[1*], Edward R Dougherty[1,2]

From Seventh Annual MCBIOS Conference. Bioinformatics: Systems, Biology, Informatics and Computation Jonesboro, AR, USA. 19-20 February 2010

Abstract

Background: Finding reliable gene markers for accurate disease classification is very challenging due to a number of reasons, including the small sample size of typical clinical data, high noise in gene expression measurements, and the heterogeneity across patients. In fact, gene markers identified in independent studies often do not coincide with each other, suggesting that many of the predicted markers may have no biological significance and may be simply artifacts of the analyzed dataset. To find more reliable and reproducible diagnostic markers, several studies proposed to analyze the gene expression data at the level of groups of functionally related genes, such as pathways. Studies have shown that pathway markers tend to be more robust and yield more accurate classification results. One practical problem of the pathway-based approach is the limited coverage of genes by currently known pathways. As a result, potentially important genes that play critical roles in cancer development may be excluded. To overcome this problem, we propose a novel method for identifying reliable subnetwork markers in a human protein-protein interaction (PPI) network.

Results: In this method, we overlay the gene expression data with the PPI network and look for the most discriminative linear paths that consist of discriminative genes that are highly correlated to each other. The overlapping linear paths are then optimally combined into subnetworks that can potentially serve as effective diagnostic markers. We tested our method on two independent large-scale breast cancer datasets and compared the effectiveness and reproducibility of the identified subnetwork markers with gene-based and pathway-based markers. We also compared the proposed method with an existing subnetwork-based method.

Conclusions: The proposed method can efficiently find reliable subnetwork markers that outperform the gene-based and pathway-based markers in terms of discriminative power, reproducibility and classification performance. Subnetwork markers found by our method are highly enriched in common GO terms, and they can more accurately classify breast cancer metastasis compared to markers found by a previous method.

Background

Given the high-throughput genomic data from microarray experiments, one challenge is to find effective biomarkers associated with a complex disease, such as cancer. Extensive work has been done to identify differentially expressed genes across different phenotypes [1-5], which can be used as diagnostic markers for classifying different disease states or predicting the clinical outcomes [6-11]. However, finding reliable gene markers is very challenging for a number of reasons. The small sample size of typical clinical data is one important factor that makes this problem difficult. We often have to select a small number of gene markers from thousands of genes based on a limited number of samples, which makes the performance of the traditional feature selection methods very unpredictable [12]. In addition to

* Correspondence: bjyoon@ece.tamu.edu
[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA
Full list of author information is available at the end of the article

this, the inherent measurement noise in microarray experiments and heterogeneity across samples aggravate this problem further [13-16]. Moreover, previous methods often select gene markers only based on their expression data. Therefore, it is possible that some of the selected gene markers may be functionally related, hence contain redundant information which may lead to the degradation of the overall classification performance.

To address the aforementioned problems, several studies proposed to interpret the expression data at the level of groups of functionally related genes, such as pathways derived from microarray studies [17-19], GO annotations [20], and other sources. Methods have been developed to capture the overall expression changes of a given pathway by jointly analyzing the expression levels of its member genes. For example, Guo et al. [21] used the mean or median expression level of the member genes as the pathway activity. Tomfohr et al. [22] analyzed the expression levels of genes in a pathway using singular value decomposition (SVD), and they used the eigenvector with the largest eigenvalue as the pathway activity. Lee et al. [23] estimated the pathway activity by computing the average expression level of the condition responsive genes (CORGs) within a pathway. More recently, another method has been proposed based on a simple probabilistic model, which estimates the pathway activity that contributes to the phenotype of interest based on the log likelihood ratios (LLR) of the member genes [24]. These pathway-based methods showed that pathway markers are generally more reliable compared to gene markers and that they lead to better or comparable classification performance [21-24]. The main advantage of the pathway-based methods is that they can reduce the effect of the measurement noise and that of the correlations between genes that belongs to the same pathway. Moreover, pathway markers can provide important biological insights into the underlying mechanisms that lead to different disease phenotypes. However, pathway-based methods also have some shortcomings. First, currently known pathways cover only a limited number of genes and may not include key genes with significant expression changes across different phenotypes. Besides, many pathways overlap with each other, hence the activity of such pathway markers may be highly correlated. One possible way to alleviate these problems is to directly identify such markers in a large protein-protein interaction (PPI) network. In a recently published paper [25], Chuang et al. tried to identify subnetwork markers by overlaying gene expression data on the corresponding proteins in a PPI network. They started from the so-called seed proteins in the PPI network that have high discriminative power and greedily grew subnetworks from them to maximize the mutual information between the subnetwork activity score and the class label. They showed that subnetwork markers yield more accurate classification results and have better reproducibility compared to gene markers.

In this paper, we propose a new method for identifying effective subnetwork markers from a PPI network by performing a *global* search for differentially expressed linear paths using dynamic programming. After finding the most discriminative linear paths, we combine overlapping paths into subnetworks through a greedy approach and use those subnetworks as diagnostic markers for classifying breast cancer metastasis. To test the effectiveness of our subnetwork markers, we perform cross validation experiments based on two independent breast cancer datasets. We compare the performance of our method with a gene-based method, a pathway-based method [24] and a previously proposed subnetwork-based method [25]. The results show that the proposed method finds reliable subnetwork markers that can accurately classify breast cancer metastasis. We also perform an enrichment analysis and show that the identified subnetwork markers are highly enriched with proteins that have common GO terms.

Results and discussion

Identification of subnetwork markers

We obtained two independent breast cancer datasets from the large-scale expression studies in Wang et al. [10] (referred as the USA dataset) and van't Veer et al. [9] (referred as the Netherlands dataset). The USA dataset contains 286 samples and the Netherlands dataset contains 295 samples. Metastasis had been detected for 78 patients in the Netherlands dataset and 107 patients in the USA dataset during the five-year follow-up visits after the surgery. The PPI network has been obtained from Chuang et al. [25], 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 9,263 proteins and 49,054 interactions for the USA dataset, and 8,380 proteins and 31,201 interactions for the Netherlands dataset.

Our proposed method integrates the gene expression data and the PPI data by overlaying the expression value of each gene on its corresponding protein in the PPI network. The subnetwork identification algorithm consists of the following three major steps:

Step 1: Search for highly discriminative linear paths whose member genes are closely correlated to each other

To find discriminative linear paths in the large PPI network, we define a scoring scheme that incorporates both the t-test statistics scores of the member genes and the correlation coefficient between their expression values. This scoring scheme takes a weighted sum of the t-scores of the member genes within a given path. The

weights depend on the correlation between the member genes and the parameter θ, where θ is introduced to control the trade off between the "discriminative power" of individual genes and the "correlation" between the member genes (see Methods). Based on the above scoring scheme, we developed an algorithm that searches for the top scoring linear paths that have length l and end at node g_i.

Step 2: Combine top scoring linear paths into a subnetwork

We initialize the subnetwork using the path with the highest score. As long as there exists a high scoring path that overlaps with the current subnetwork, we combine them and check if the discriminative power of the new subnetwork is larger than that of the previous subnetwork. If the discriminative power improves, we keep the new subnetwork. Otherwise, we keep the previous subnetwork and check the next best path. To evaluate the discriminative power of subnetworks, we applied the probabilistic pathway activity inference method proposed in [24] to infer the subnetwork activity. The discriminative power of a subnetwork is assessed by computing the t-test statistics score of the subnetwork activity.

Step 3: Update the PPI network

After identifying the discriminative subnetwork, we update the PPI network by removing the proteins in the identified subnetwork from the current PPI network. In order to find additional *non-overlapping* subnetworks, we repeat the search process from **Step 1**.

In order to control the size of the identified subnetworks, we restricted the length of the linear paths to be less than 8. For a given l and for every node g_i in the network, we identified the top 20 linear paths with the highest scores, whose length is l and end at the given node g_i. To construct the subnetwork marker that can be used as a diagnostic marker for breast cancer metastasis, we chose the top 100 scoring linear paths whose length are within a given range $5 \le l \le 8$. The selected linear paths were combined into a single subnetwork as described in **Step 2**. To find the best θ, we repeated the experiment for six different values θ = 1, 2, 4, 8, 16 and ∞. For every value of θ, we identified 50 subnetwork markers for each dataset using the proposed method. The statistics of the identified subnetworks for the two datasets are shown in Table 1. We can see that the overlap between the subnetwork markers identified on different datasets is around 25%, which is significantly larger than the overlap reported in Chuang et al. (12.7%) [25].

The identified subnetworks are enriched with proteins that have common GO terms

We identified 50 discriminative subnetworks using the proposed method for both the USA dataset and the Netherlands dataset (θ = 8). The identified subnetworks consist of 1035 and 976 genes, respectively. Next, we analyzed the identified subnetworks using FuncAssociate [26], which is a web application designed for characterizing large collections of genes and proteins. It performs a Fisher's Exact Test (FET) analysis to identify Gene Ontology (GO) [20] attributes that are shared by a fraction of the entries in a given set of genes or proteins. At a significance threshold of 0.01, 78% and 84% of the subnetworks that were respectively identified using the USA dataset and the Netherlands dataset were enriched with proteins that share common GO terms. These GO terms generally correspond to cell growth and death, cell proliferation and replication, cell and tissue remodeling, circulation and coagulation, or metabolism. Examples of the identified subnetworks are shown in Figure 1, where we can see that the proposed method is capable of finding subnetwork markers that also include genes that are oppositely regulated. The enrichment analysis results of the sample subnetworks obtained using FuncAssociate are shown in Table 2.

The subnetwork markers identified by the proposed method are more discriminative and reproducible

We first evaluated the subnetwork markers identified using the proposed method. For a given θ, we identified the subnetwork markers based on one dataset and estimated their discriminative power on the same dataset. The discriminative power of the subnetwork marker was estimated as the absolute t-test statistics score of the subnetwork activity. Subnetwork markers were then sorted in the decreasing order of t-score. Next, to show the reproducibility of our subnetwork markers, we identified the top 50 markers based on one dataset and evaluated their discriminative power on the other dataset. Again, subnetwork markers were sorted according to their discriminative power. Figure 2 shows the discriminative power of subnetwork markers identified using six different values of θ, where the x-axis corresponds to the top K markers being considered, and the y-axis shows the mean absolute t-score of the top K markers (K = 10, 20, 30, 40, 50). Figure 2A and Figure 2B show the results obtained from the USA dataset and the Netherlands dataset, respectively. Figure 2C shows the discriminative power of the subnetwork markers selected based on the Netherlands dataset and evaluated using the USA dataset. Figure 2D shows the discriminative power of the markers selected based on the USA dataset and evaluated using the Netherlands dataset. As we can see from these results, the discriminative power of the identified subnetwork markers is not very sensitive to the choice of θ. To further compare the identified subnetwork markers with other markers, we used θ = 8 which showed good performance in average.

Table 1 Statistics of the subnetwork markers identified by the proposed method

θ		Size		Number of genes	Number of genes in common
		mean	standard deviation		
1	USA	16.8	10.17	840	213
	Netherlands	14.62	8.69	731	
2	USA	18.22	12.3	911	233
	Netherlands	16	10.34	801	
4	USA	18	12.8	901	202
	Netherlands	17.28	11.4	864	
8	USA	20.7	13.38	1035	252
	Netherlands	19.52	12.57	976	
16	USA	20.2	11.13	1010	201
	Netherlands	16.64	10.89	832	
∞	USA	22.32	14.86	1116	266
	Netherlands	21.92	10.67	1096	

For each θ, we show the mean and standard deviation of the subnetwork size as well as the total number of genes covered by the identified subnetworks. We also show the number of genes shared by the subnetworks identified using the respective breast cancer datasets.

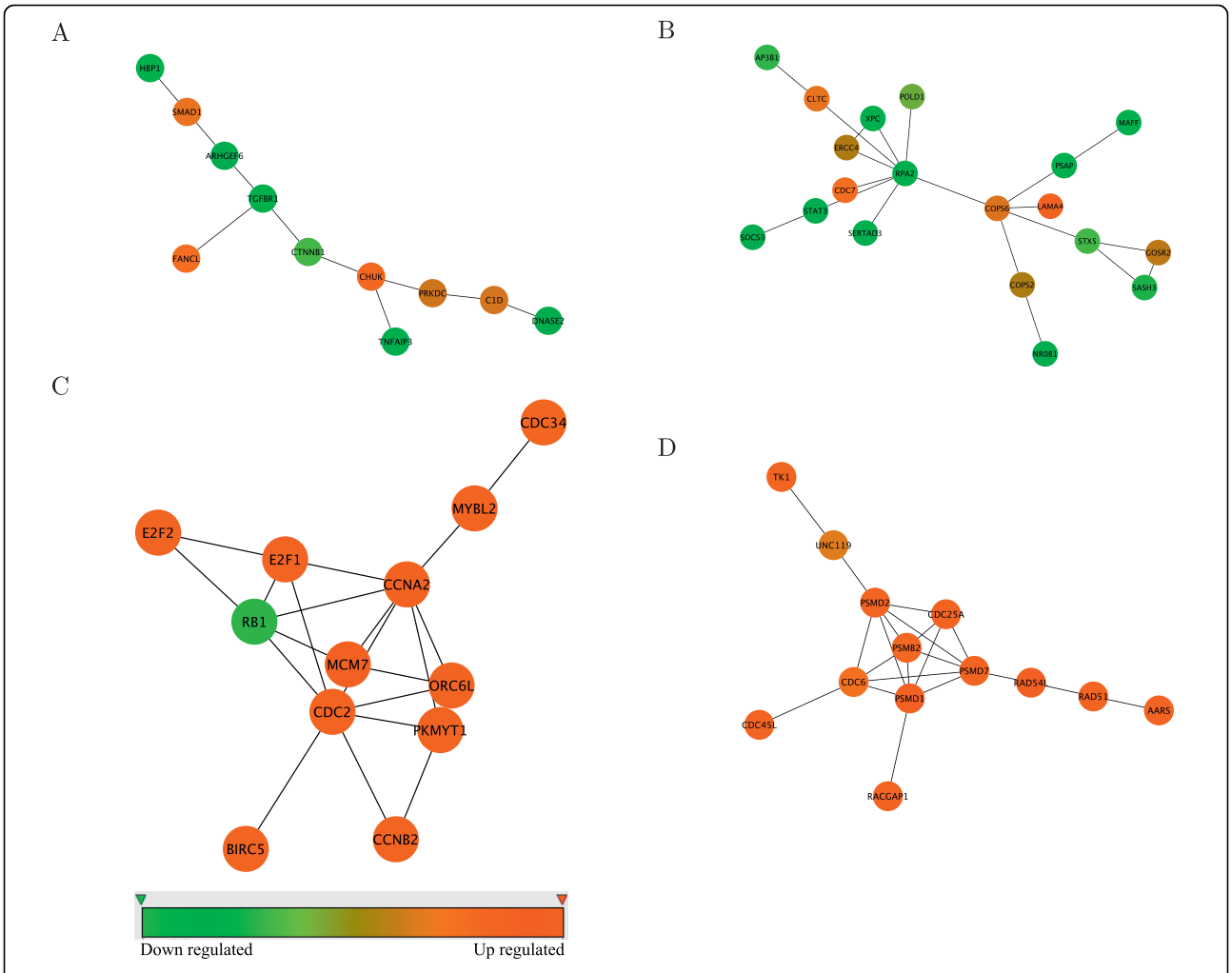

Figure 1 Sample subnetworks identified using the proposed method. (A), (B) are examples of subnetworks identified using the USA dataset. (C), (D) are examples of subnetworks identified using the Netherlands dataset. Red (green) implies that the gene is upregulated (downregulated) in breast cancer samples with metastasis.

Table 2 Enrichment analysis results for the sample subnetworks shown in Figure 1

Subnetwork	Attribute ID	*P*-value	Attribute name
A	GO:0045165	0.024	cell fate commitment
	GO:0012501	0.001	programmed cell death
	GO:0008219	0.006	cell death
	GO:0016265	0.006	Death
	GO:0006915	0.017	Apoptosis
B	GO:0000718	< 0.001	nucleotide-excision repair, DNA damage removal
	GO:0006308	0.046	DNA catabolic process
	GO:0043566	0.040	structure-specific DNA binding
C	GO:0051318	0.039	G1 phase
	GO:0022403	< 0.001	cell cycle phase
	GO:0005654	< 0.001	Nucleoplasm
	GO:0000280	0.009	nuclear division
	GO:0007067	0.009	Mitosis
	GO:0048285	0.009	organelle fission
	GO:0051301	0.001	cell division
	GO:0022402	< 0.001	cell cycle process
	GO:0007049	0.000	cell cycle
	GO:0051726	0.008	regulation of cell cycle
	GO:0044428	0.001	nuclear part
	GO:0005634	0.024	Nucleus
D	GO:0005838	< 0.001	proteasome regulatory particle
	GO:0000076	0.016	DNA replication checkpoint
	GO:0032297	0.016	negative regulation of DNA replication initiation
	GO:0030174	0.030	regulation of DNA replication initiation
	GO:0000502	0.010	proteasome complex
	GO:0031145	< 0.001	anaphase-promoting complex-dependent proteasomal ubiquitin-dependentprotein catabolic process
	GO:0051436	< 0.001	negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle
	GO:0051352	< 0.001	negative regulation of ligase activity
	GO:0051437	< 0.001	positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle
	GO:0051444	< 0.001	negative regulation of ubiquitin-protein ligase activity
	GO:0051443	0.001	positive regulation of ubiquitin-protein ligase activity
	GO:0051439	0.001	regulation of ubiquitin-protein ligase activity during mitotic cell cycle
	GO:0051351	0.001	positive regulation of ligase activity
	GO:0051438	0.002	regulation of ubiquitin-protein ligase activity
	GO:0051340	0.002	regulation of ligase activity
	GO:0010498	0.007	proteasomal protein catabolic process
	GO:0043161	0.007	proteasomal ubiquitin-dependent protein catabolic process
	GO:0022402	< 0.001	cell cycle process
	GO:0006511	0.050	ubiquitin-dependent protein catabolic process
	GO:0006259	0.050	DNA metabolic process

Next, we compared the identified subnetwork markers with gene markers, pathways markers [24] and the subnetwork markers identified by Chuang et al. [25]. For gene markers, we selected the top 50 genes based on the absolute *t*-score among all genes covered by the 50 identified subnetworks. For pathway markers, we selected the top 50 pathways among the 639 pathways in the C2 curated gene sets in MsigDB (Molecular Signatures Database) [17]. We also obtained the subnetworks identified by Chuang et al. [25] from the Cell Circuits database [27] (149 discriminative subnetworks for the Netherlands dataset and 243 subnetworks for the USA dataset). We chose the top 50 subnetworks out of 149 subnetworks based on the Netherlands dataset and the top 50 subnetworks out of 243 subnetworks based on the USA dataset. The pathways and subnetworks were ranked using the scheme proposed by Tian et al. [18], based on the average absolute *t*-test statistics score of all the member genes. For subnetwork markers identified by Chuang et al., we computed the *t*-scores of

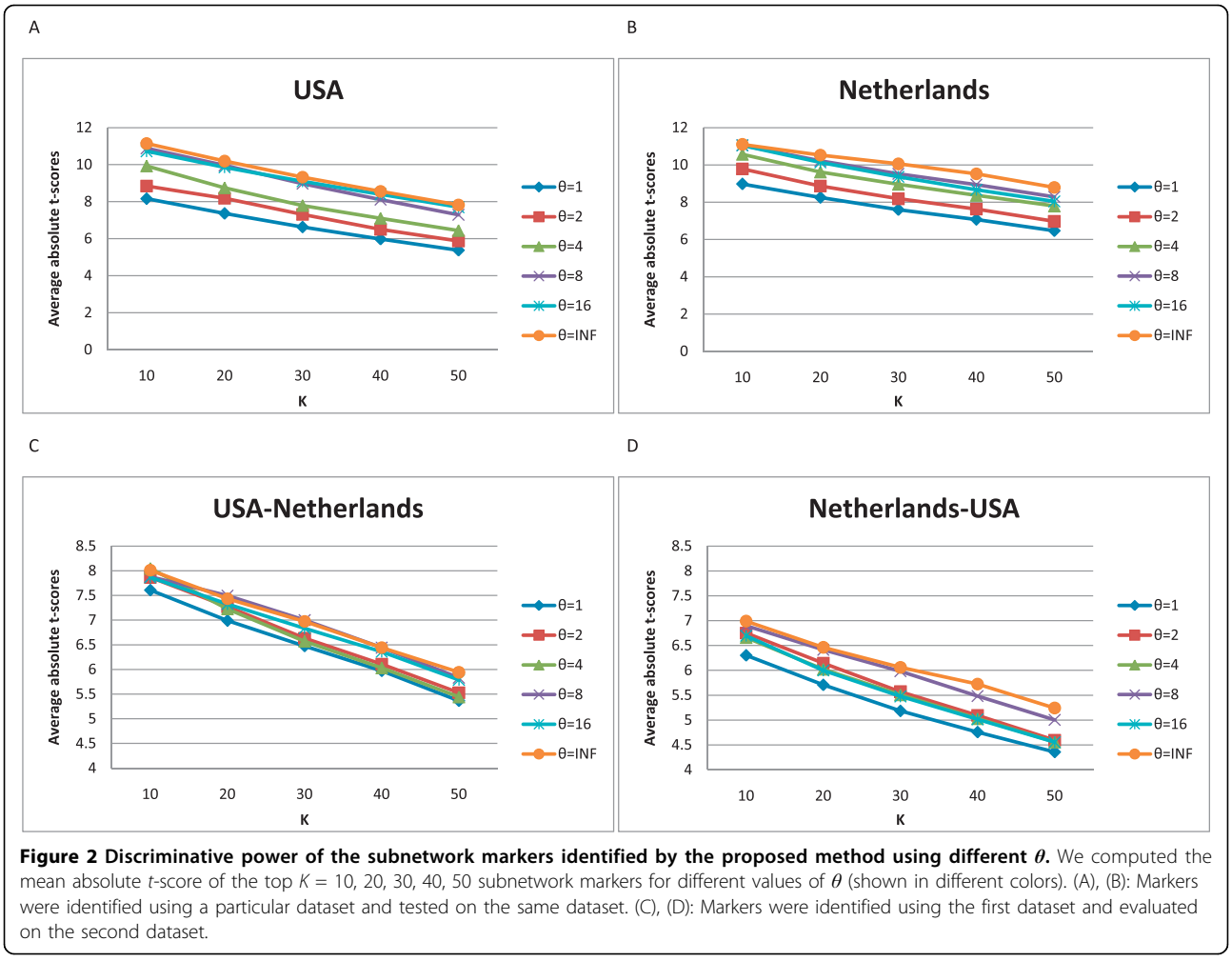

Figure 2 Discriminative power of the subnetwork markers identified by the proposed method using different θ. We computed the mean absolute *t*-score of the top K = 10, 20, 30, 40, 50 subnetwork markers for different values of θ (shown in different colors). (A), (B): Markers were identified using a particular dataset and tested on the same dataset. (C), (D): Markers were identified using the first dataset and evaluated on the second dataset.

their member genes using the original expression values. For pathway markers, *t*-scores of the member genes were computed using their log-likelihood ratios as in [24] (see Methods). To assess the discriminative power of the subnetwork markers identified using the proposed method, their activity score was inferred using the probabilistic inference method proposed in [24]. For subnetwork markers identified by Chuang et al., we inferred their activity score using the mean expression value of the member genes as reported in their paper [25].

The discriminative power of these different markers are shown in Figure 3. As we can see in Figure 3, subnetwork markers identified by our method are more discriminative compared to other markers. Moreover, it can be seen that they also retain higher discriminative power across different datasets.

Subnetwork markers identified by the proposed method improve classification performance

To evaluate the performance of the classifiers that are constructed using the subnetwork markers identified by the proposed method, we performed the following within-dataset and cross-dataset cross-validation experiments.

In the 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 ten folds of equal size, one of them was withheld as the "test set" and the remaining nine were used for training the classifier. In the training set, six folds (referred as the "marker ranking set") were used to rank the subnetwork markers according to their discriminative power and to build the classifier using logistic regression. The other three folds (referred as the "feature selection set") were used for feature selection. We started with the top ranked subnetwork marker and enlarged the feature set by adding features sequentially. Every time we included a new subnetwork marker into the feature set, a new classifier was built using the marker ranking set and it was tested on the feature selection set. For all the samples in the feature selection set, the classifier can compute the posterior probabilities of the

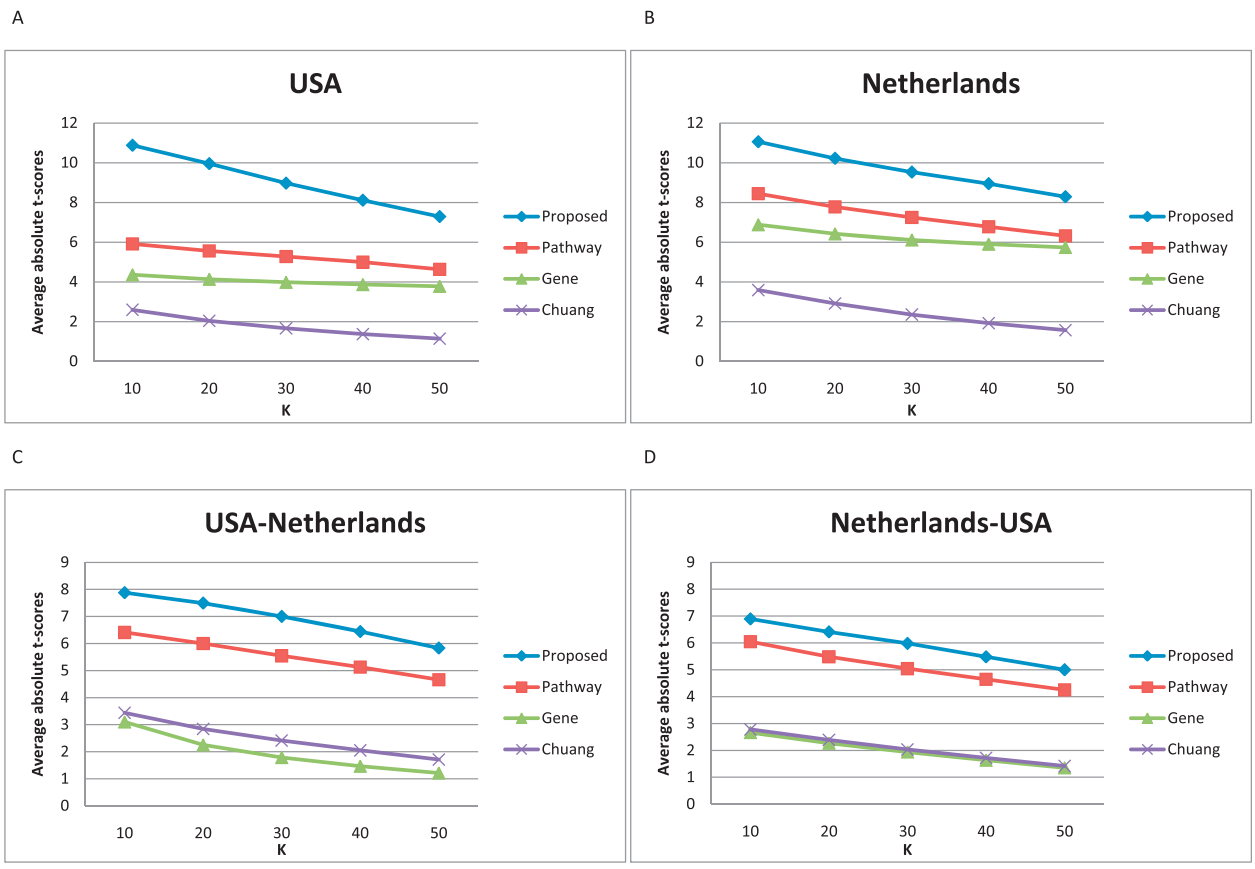

Figure 3 Discriminative power of different types of markers. We evaluated the discriminative power of the subnetwork markers identified using the proposed method, and compared them with gene markers, pathway markers [24], and the subnetwork markers identified by Chuang et al. [25]. Mean absolute t-score is shown for the top K = 10, 20, 30, 40, 50 markers. (A), (B): Markers were identified using a particular dataset and tested on the same dataset. (C), (D): Markers were identified using the first dataset and evaluated based on the second dataset.

class label (metastasis versus metastasis-free), based on which we can estimate the AUC (Area Under ROC Curve) [28]. The AUC metric provides a useful statistical summary of the classification performance over the entire range of sensitivity and specificity. We retained the new subnetwork marker if the AUC (estimated on the feature selection set) increased; otherwise, we discarded the subnetwork marker and continued to test the remaining ones. The above experiment was repeated 500 times based on 50 random ten-fold splits. The average AUC was reported as the classification performance measure.

To evaluate the reproducibility of the subnetwork markers, we performed the following cross-dataset experiments. We first identified the top 50 subnetwork markers based on one dataset and performed cross-validation experiments on the other dataset, following a similar procedure that was used in the previously described within-dataset experiments.

For comparison, we also performed similar within-dataset and cross-dataset experiments using gene markers, pathway markers and the subnetwork markers identified by Chuang et al., respectively. For each method, we limited the feature set to the top 50 markers for each dataset. Figure 4 shows the classification performance based on the subnetwork markers identified by the proposed method for different values of θ. We found that the AUC for both within-dataset and cross-dataset experiments first increases with increasing θ and starts to drop after certain point. At θ = 8, the AUC values for both cross-dataset experiments are relatively larger than those at other values of θ. Also, the AUC values for both within-dataset experiments at θ = 8 compare favorably with those at different θ, which implies that the trade off between maximizing the discriminative power and increasing the correlations of the member genes is well balanced.

To compare the classification performance of the identified subnetwork markers with other types of markers, we set θ = 8. Based on this setting, we compared our subnetwork markers with gene markers, pathway markers and the subnetwork markers from Chuang

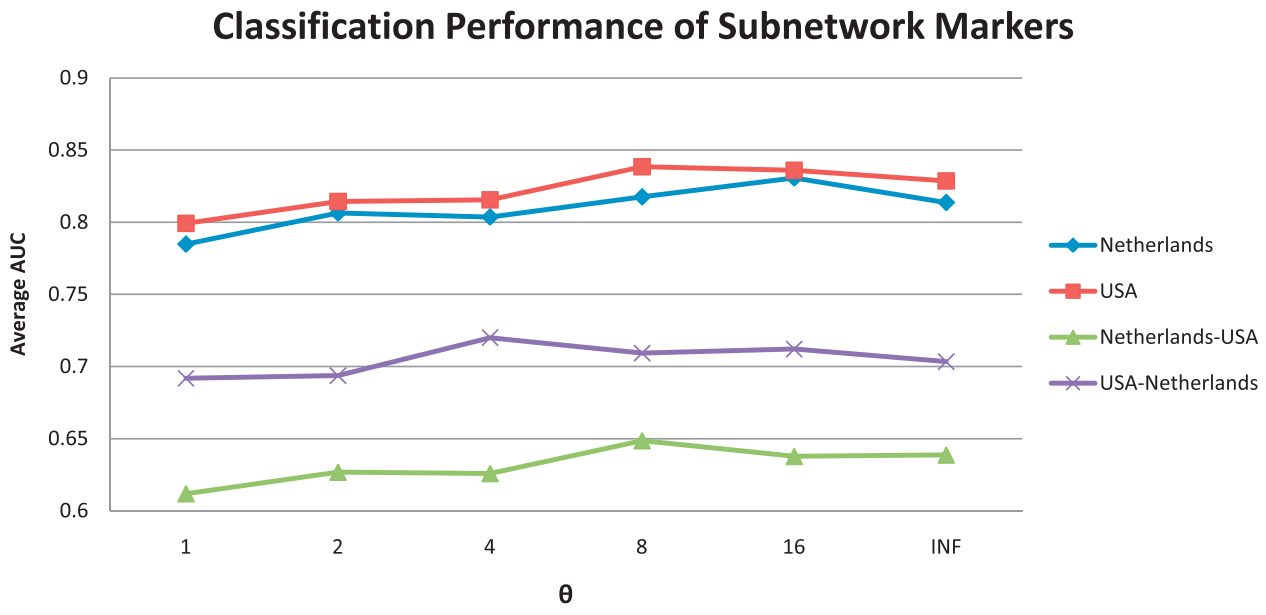

Figure 4 Classification performance of the identified subnetwork markers for different θ. The line plots show the average AUC for classifiers based on subnetwork markers identified using θ = 1, 2, 4, 8, 16, ∞. The legends USA, Netherlands denote the results of within-dataset experiments based on the USA dataset and the Netherlands dataset, respectively. The legends USA-Netherlands, Netherlands-USA denote the results of cross-dataset experiments where markers were identified based on the first dataset and tested based on the second dataset.

et al. using the experimental designs described above. Figure 5 summarizes the classification performance of the proposed approach, in comparison with the other methods. The two bar charts on the left of Figure 5 show the AUC of the within-dataset experiments. As shown in Figure 5, classifiers based on the subnetwork markers identified by the proposed method perform significantly better than the classifiers based on other types

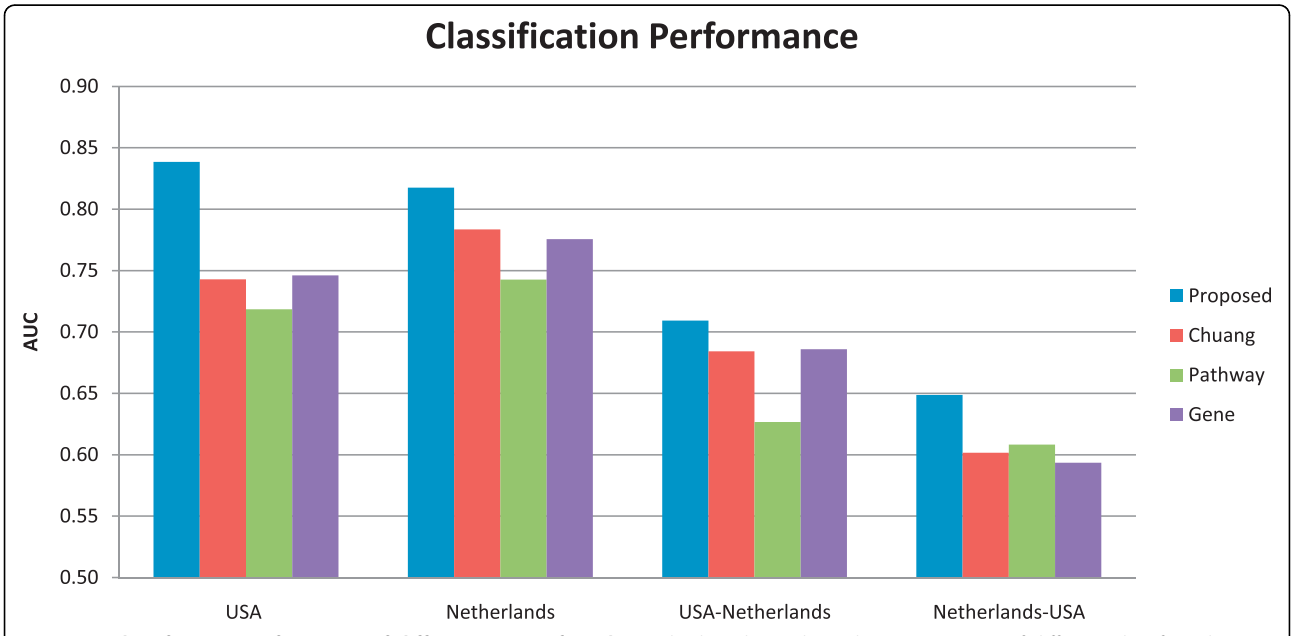

Figure 5 Classification performance of different types of markers. The bar charts show the average AUC of different classifiers that use subnetwork markers identified by the proposed method, gene markers, pathway markers, and subnetwork markers found by Chuang et al.'s method. Results of the within-dataset experiments based on the USA and Netherlands dataset are shown in the two bar charts on the left. The two bar charts on the right show the results of the cross-dataset experiments, where markers were identified based on the first dataset and tested based on the second dataset.

of markers. The results of the cross-dataset experiments are shown in the two bar charts on the right of Figure 5. Again, we can see that the classifiers built on the subnetwork markers predicted by our method significantly outperform those based on other markers. This indicates that the predicted subnetwork markers are more reproducible compared to other markers.

Figure 6 shows the classification error of the classifiers built using different types of markers at different TPR (true positive rate). As shown in Figure 6, the error curve that corresponds to the proposed markers always lies below others, which implies that classifiers built on our subnetwork markers yield a lower error rate at any fixed sensitivity level.

Conclusions

In this paper, we proposed a new method for identifying effective subnetwork markers in a protein-protein interaction (PPI) network. As shown throughout this paper, integrating the PPI network with microarray data can overcome some of the shortcomings of the gene-based and pathway-based methods. First of all, using a genome-scale PPI network provides a better coverage of the genes in the microarray studies compared to using known pathways obtained from public databases. Second, the network topology provides prior information about the relationship between proteins, hence the genes that code for these proteins. Subnetworks identified by integrating the network structure and the gene

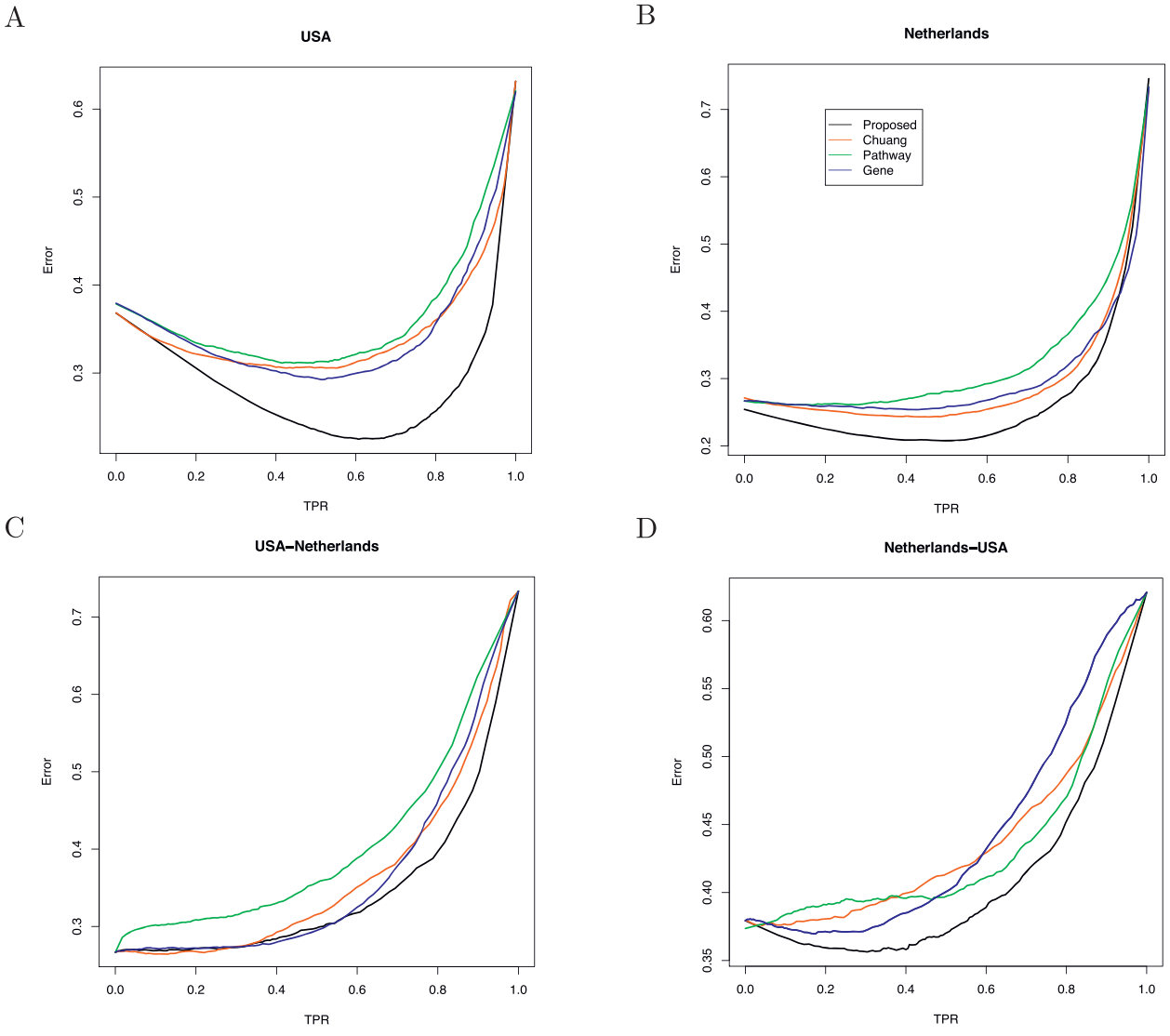

Figure 6 Classification error at different TPR (true positive rate) for different types of markers. (A), (B) show the results of the within-dataset experiments based on the USA dataset and the Netherlands dataset, respectively. (C), (D) show the results of the cross-dataset experiments, where markers were identified using the first dataset and tested based on the second dataset.

expression data can cluster proteins (or genes) that are functionally related to each other. By aggregating the expression values of the member genes, subnetwork markers can avoid selecting single gene markers with redundant information. Furthermore, the discriminative subnetworks identified by the proposed method can also provide us with important clues about the biological mechanisms that lead to different disease phenotypes. The proposed method finds top scoring linear paths using dynamic programming and combines them into a subnetwork by greedily optimizing the discriminative power of the resulting subnetwork marker. We developed a scoring scheme that is used by the search algorithm to find linear paths that consist of discriminative genes that are highly correlated to each other. The proposed algorithm allows us to control the trade off between maximizing the discriminative power of the member genes within a given linear path and increasing the correlation between the member genes, by choosing the appropriate value for θ. As the subnetwork markers are constructed based on the top scoring linear paths, instead of single genes, the proposed method is expected to yield more robust subnetwork markers. Another important advantage of our method is that it can find non-overlapping subnetwork markers. This can reduce the overall redundancy among the identified markers. In this paper, the activity of the identified subnetwork markers were inferred using the probabilistic activity inference scheme proposed in [24]. This allows us to find better subnetwork markers, since it can assess their discriminative power more effectively.

As shown in this paper, the identified subnetwork markers consist of proteins that share common GO terms. The classifiers based on the subnetwork markers identified using the proposed method were shown to achieve higher classification accuracy in both within-dataset and cross-dataset experiments compared to classifiers based on other markers. These results suggest that the method proposed in this paper can find effective subnetwork markers that can more accurately classify breast cancer metastasis and are more reproducible across independent datasets.

Methods

Overview

Given a large PPI network, we want to find subnetwork markers whose activity is highly indicative of the disease state of interest. For this purpose, we first need a method for inferring the activity of a given subnetwork and evaluating its discriminative power. There exist different ways for computing the activity score of a given group of genes [24]. Recently, we proposed a probabilistic pathway activity inference scheme, which was shown to outperform 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 simple scheme proposed in Tian et al. [18]. This scheme scores a pathway marker by computing the average absolute t-score of its member genes. It has been shown to be effective in evaluating the discriminative power of pathway markers in [24]. Since our goal is to find groups of genes that display coordinated expression patterns, we modified Tian et al.'s scoring scheme to incorporate the correlation between the genes within a given pathway. This new method scores a given pathway by taking the weighted sum of the absolute t-scores of its member genes, where the weights are computed using the correlation coefficients between the member genes. The general outline of the proposed algorithm is as follows. Based on the above scoring scheme, we first search for differentially expressed linear paths in the PPI network. Then, the top paths that overlap with each other are greedily combined into a subnetwork by maximizing the discriminative power of the resulting subnetwork, evaluated by the method proposed in [24]. 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. 7.

Probabilistic inference of subnetwork activity

Here we provide a brief review of the probabilistic activity inference method proposed in [24]. Suppose we have a subnetwork G_s that consists of n proteins which correspond to n different genes $\{g_1, g_2,...,g_n\}$. Assume that the expression level x_i of a gene g_i follows the distribution

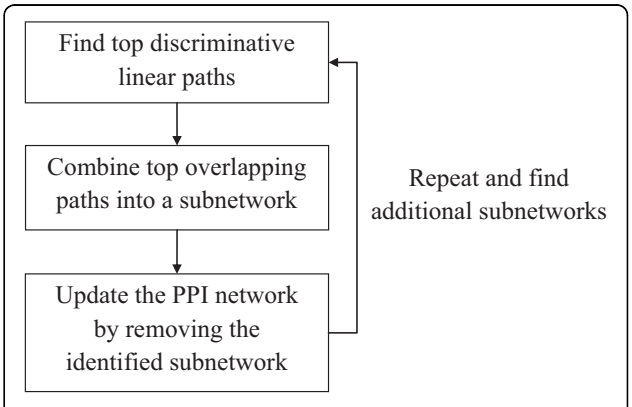

Figure 7 Illustration of the proposed method.

$f_i^k(x_i)$ under phenotype k = 1, 2. The log-likelihood ratio (LLR) [24] between the two phenotypes is computed as follows

$$\alpha(x_i)=\log(f_i^1(x_i)/f_i^2(x_i)).$$

In order to estimate the conditional probability density function $f_i^k(x_i)$, we assume that the gene expression level of gene g_i under phenotype k follows a Gaussian distribution with mean μ_i^k and standard deviation σ_i^k. The parameters are empirically estimated using the samples with phenotype k. Given the log-likelihood ratio of each gene, the subnetwork activity A_{G_s} is defined as the sum of the log-likelihood ratios of the member genes $A_{G_s}=\sum_{i=1}^{n}\alpha(x_i)$.

Evaluating the discriminative power of linear paths in the PPI network

A linear path $\lambda = \{g_1, g_2,...,g_n\}$ in a given PPI network G is defined as a group of genes, where the proteins that correspond to g_i and g_{i+1} are connected for i = 1,..., n – 1. To evaluate the discriminative power of a linear path, we first evaluate the discriminative power of each gene g_i by computing the t-test statistics score of the log-likelihood ratio $\alpha(x_i)$, denoted as $t_\alpha(g_i)$. Then, we compute the Pearson product-moment correlation coefficient to measure the correlation between the log-likelihood ratios of $\forall g_i, g_j \in \lambda$. The correlation matrix is given by

$$\sum(\lambda)=\begin{bmatrix}1 & \rho_{12} & \cdots & \rho_{1n}\\ \rho_{21} & 1 & \cdots & \rho_{2n}\\ \vdots & \vdots & & \vdots\\ \rho_{n1} & \rho_{n2} & \cdots & 1\end{bmatrix}$$

where ρ_{ij}, $i \neq j$ is the correlation coefficient between the log-likelihood ratios of g_i and g_j. The score of the pathway λ is defined as following

$$S(\lambda)=\frac{1}{n^2}\left[t_\alpha(g_1),t_\alpha(g_2),.....,t_\alpha(g_n)\right]\cdot\Sigma'(\lambda)\cdot J$$

where $\Sigma'(\lambda)=\frac{1}{1+\theta}\left[\left(\Sigma'(\lambda)-I\right)+\theta\cdot I\right]$ (I is the identity matrix), and J is an all-one-column vector. We use a normalization factor of $\frac{1}{n^2}$ to ensure that the overall score does not depend on the length of the path. We use θ to control the trade off between maximizing the discriminative power of the genes within the identified path and increasing the correlation between its member genes. When θ = 0, the weight for the t-score of a given gene g_i is determined by the average correlation between the log-likelihood ratios of g_i and g_j, where $j \neq i$. As θ increases, we give more weight on the discriminative power of individual genes than the correlation between member genes. Especially, when $\theta \to \infty$, we get $\Sigma'(\lambda) = I$. In this case, the pathway score $S(\lambda)$ is simply the average t-score of the member genes in λ, and the proposed subnetwork marker identification method reduces to its preliminary version proposed in [29]. The above scoring scheme is used for finding the top linear paths in the network G as we describe in the following section.

Searching for discriminative linear paths

Let $G = (E, V)$ 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 G. Then 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 scoring scheme defined in the previous section, we search for top discriminative linear paths using dynamic programming. We define $\lambda(v_i, l)$ as the optimal linear path among all linear paths that have length l and end at v_i. The score of this optimal path is defined as

$s(v_i, l) = t_\alpha[\lambda(v_i,l)]\ \Sigma'[\lambda(v_i, l)]$.

Here, only paths 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 $\forall v_i \in V$,

$$s(v_i,l)=\max_{v_j}\{t(\lambda(v_j,l),v_i)\cdot\Sigma'(\lambda(v_j,l),v_i)+\log(E[v_i,v_j])\},$$

$$v_j^*=\arg\max_{v_j}\{t(\lambda(v_j,l),v_i)\cdot\Sigma'(\lambda(v_j,l),v_i)+\log(E[v_i,v_j])\},$$

if $s(v_i, l) > 0$, then

$$\lambda(v_i,l) = \lambda\left(v_j^*,l-1\right)\cup\{v_i\}.$$

end

end

(iii) **Termination:**

for $\forall v_i \in V$, $1 \leq l \leq L$,

$S(\lambda(v_i, l)) = s(v_i, l)/l^2$.(1)

Although the above algorithm finds only the top path for every (v_i, l), we can easily modify it to find the top M discriminative paths. Increasing M allows us to find better linear paths with higher discriminative power, but it will also increase the computational complexity of the algorithm.

Combining top overlapping paths into a subnetwork

Based on (1), we choose the m top scoring paths $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ whose length is within a given range $[L_{\min}, L_{\max}]$. Next, the paths in Λ are combined into a subnetwork G_s so that its discriminative power $R(G_s)$ is locally optimized. This process is carried out as follows:

(i) $G_s \leftarrow \lambda_i$, $G_{temp} \leftarrow G_s$, $i = 1$.

(ii) $i = i + 1$; If $\lambda_i \cap G_s \neq \emptyset$, $G_{temp} \leftarrow G_{temp} \cup \lambda_i$.

(iii) If $R(G_{temp}) > (1 + \in)R(G_s)$, $G_s \Re G_{temp}$; else $G_{temp} \Re G_s$.

(iv) Go to (ii) if $i < m$; otherwise, terminate.

Here $\in$ is set as 0.01 to avoid over-fitting to the expression data. We used the activity inference method in [24] to computed the actual activity score of G_s. Then, $R(G_s)$ is computed as the t-test statistics of the subnetwork activity score.

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

Acknowledgements

We would like to thank the authors of [25], especially H.Y. Chuang and T. Ideker, for sharing the PPI network and their helpful communication.
This article has been published as part of *BMC Bioinformatics* Volume 11 Supplement 6, 2010: Proceedings of the Seventh Annual MCBIOS Conference. Bioinformatics: Systems, Biology, Informatics and Computation. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/11?issue=S6.

Author details

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA. [2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA.

Authors contributions

Conceived and designed the experiments: JS BJY ERD. Performed the experiments: JS. Analyzed the data: JS BJY ERD. Wrote the paper: JS BJY ERD.

Competing interests

The authors declare that they have no competing interests.

Published: 7 October 2010

References

1. Efron B, Tibshirani R: **Empirical bayes methods and false discovery rates for microarrays.** *Genet. Epidemiol* 2002, **23**:70-86.
2. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17**:509-519.
3. Kepler TB, Crosby L, Morgan KT: **Normalization and analysis of DNA microarray data by self-consistency and local regression.** *Genome Biol.* 2002, **3, RESEARCH0037.**
4. Ideker T, Thorsson V, Siegel AF, Hood LE: **Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data.** *J. Comput. Biol.* 2000, **7**:805-817.
5. Chen Y, Dougherty ER, Bittner ML: **Ratio-based decisions and the quantitative analysis of cDNA microarray images.** *Journal of Biomedical Optics* 1997, **2**:364-374.
6. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**:503-511.
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
8. Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nat. Genet.* 2003, **33**:49-54.
9. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**:530-536.
10. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365**:671-679.
11. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc. Natl. Acad. Sci. U.S. A.* 2001, **98**:11462-11467.
12. Hua JDE, Tembe WD: **Performance of feature-selection methods in the classification of high-dimension data.** *Pattern Recognition* 2008, **42**:409-424.
13. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21**:171-178.
14. Symmans WF, Liu J, Knowles DM, Inghirami G: **Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions.** *Hum. Pathol.* 1995, **26**:210-216.
15. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM: **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.** *Science* 2005, **310**:644-648.
16. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat. Genet.* 2003, **34**:267-273.
17. Subramanian A, Tamayo P, feature VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc. Natl. Acad. Sci. U.S.A.* 2005, **102**:15545-15550.
18. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc. Natl. Acad. Sci. U.S.A.* 2005, **102**:13544-13549.
19. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA, Marks JR, Dressman HK, West M, Nevins JR: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439**:353-357.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat. Genet.* 2000, **25**:25-29.
21. Guo Z, Zhang T, Li X, Wang Q, Xu J, Yu H, Zhu J, Wang H, Wang C, Topol EJ, Wang Q, Rao S: **Towards precise classification of cancers based on robust gene functional expression profiles.** *BMC Bioinformatics* 2005, **6**:58.
22. Tomfohr J, Lu J, Kepler TB: **Pathway level analysis of gene expression using singular value decomposition.** *BMC Bioinformatics* 2005, **6**:225.
23. Lee E, Chuang HY, Kim JW, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification.** *PLoS Comput. Biol.* 2008, **4**:e1000217.
24. Su J, Yoon BJ, Dougherty ER: **Accurate and reliable cancer classification based on probabilistic inference of pathway activity.** *PLoS ONE* 2009, **4**: e8161.
25. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol. Syst. Biol.* 2007, **3**:140.
26. Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP: **Next generation software for functional trend analysis.** *Bioinformatics* 2009, **25**:3043-3044.

27. Mak HC, Daly M, Gruebel B, Ideker T: **CellCircuits: a database of protein network models.** *Nucleic Acids Res.* 2007, **35**:D538-545.
28. Fawcett T: **An introduction to ROC analysis.** *Patt Recog Letters* 2006, **27**:861-874.
29. Su J, Yoon BJ: **Identifying reliable subnetwork markers in protein-protein interaction network for classification of breast cancer metastasis.** *Acoustics, Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* 2010, 525-528 [http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5495633&isnumber=5494886].

doi:10.1186/1471-2105-11-S6-S8
Cite this article as: Su *et al.*: Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC Bioinformatics* 2010 **11**(Suppl 6):S8.