BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

A randomized steiner tree approach for biomarker discovery and classification of breast cancer metastasis

Md. Jamiul Jahid [1] and Jianhua Ruan [1]

[1]Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: DNA microarray has become an important tool to help identify biomarker genes for improving the prognosis of metastatic breast cancer - a leading cause of cancer-related deaths in women worldwide. Recently, pathway-level relationships between genes have been increasingly used to build more robust classification models which also can provide useful biological insights. Due to the unavailability of complete pathways, protein-protein interaction (PPI) network is becoming more popular to researcher and opens a way to investigate the developmental process of breast cancer. Here, a network-based method is proposed to combine microarray gene expression profiles and PPI network for biomarker discovery for breast cancer metastasis. The key idea is to identify a small number of genes to connect differentially expressed genes into a single component in a PPI network; these intermediate genes contain important information about the pathways involved in metastasis and have a high probability of being biomarkers.

Results: We applied this approach on three breast cancer microarray datasets, and identified significant numbers of well-known biomarker genes for breast cancer metastasis. Those selected genes are enriched with biological processes and pathways related to carcinogenic process, and, importantly, have higher stability across different datasets than in previous studies. Furthermore, those genes significantly increased cross-data classification accuracy in breast cancer metastasis. The randomized Steiner tree based approach described here is a new way to discover biomarker genes for breast cancer, and improves the prediction accuracy of metastasis. The analysis is limited here only to breast cancer, but can be easily applied to other diseases.

Contact: {mjahid,jruan}@cs.utsa.edu

1 INTRODUCTION

The identification of marker genes involved in cancer is a central problem in system biology. Many studies have used gene expression data for marker identification in breast cancer and other diseases (Golub *et al.*, 1999; Sotiriou and Pusztai, 2009). However, noisy data, small sample sizes, and heterogeneous experimental platforms make the marker selection procedure difficult and dataset-specific. As a result, different studies on the same disease often have very few gene markers in common. For example, two studies (van 't Veer *et al.*, 2002; Wang *et al.*, 2005) identified 70 and 76 gene marker for breast cancer, which were also validated later by two other studies (van de Vijver *et al.*, 2002; Desmedt *et al.*, 2007), but they have only three genes in common.

To improve the stability of marker selection, other complementary genomic information such as pathways has been used (Pavlidis *et al.*, 2004; Tian *et al.*, 2005; Wei and Li, 2007). The problem of pathway-based approach, however, is that the majority of human genes are not assigned to a specific pathway (Chuang *et al.*, 2007); therefore there is a strong possibility that a true marker may be out of consideration for not being assigned to a pathway. To circumvent this problem, (Chuang *et al.*, 2007) proposed to incorporate protein-protein interaction (PPI) networks for discovering small sub-networks, which may represent novel pathways, as potential markers. They found that such subnetwork-based markers can both improve classification accuracy and increase cross-dataset stability. Other studies have attempted to use gene co-expression networks or hybrid networks developed from various sources instead of PPI networks (Ma *et al.*, 2010; Wu *et al.*, 2010). Recently several studies also paid much attention to the association between PPI network topology and disease. For example, Taylor *et al.* (2009) found that inter-modular hubs are more associated with breast cancer than intra-modular hubs; Dezso *et al.* (2009) used pair-wise shortest paths between differentially expressed genes to identify candidate markers.

In this study we propose a network topology-based approach to identify candidate biomarker genes, motivated by the key observation that disease genes play a role in connecting differentially expressed (DE) genes in PPI networks (Chuang *et al.*, 2007). The main idea of our approach is to find a small number of genes that can connect DE genes into a singly connected component in a PPI network, which maps to the well-known Steiner tree problem and is solved using a heuristic algorithm. In addition, we combine multiple suboptimal Steiner trees to increase the chance of finding the optimal solution and to capture alternative pathways. Finally, we propose a method to rank the candidate markers by their topology in the PPI network. Applying our approach on three breast cancer datasets, we found that the candidate markers selected by our method are highly enriched in pathways that are well-known to be dysregulated in breast cancer metastasis, and cover a significant number of known breast cancer susceptibility genes. Remarkably, the markers identified from multiple datasets have much higher reproducibility than in previous studies, and significantly increase the cross-datasets classification accuracy.

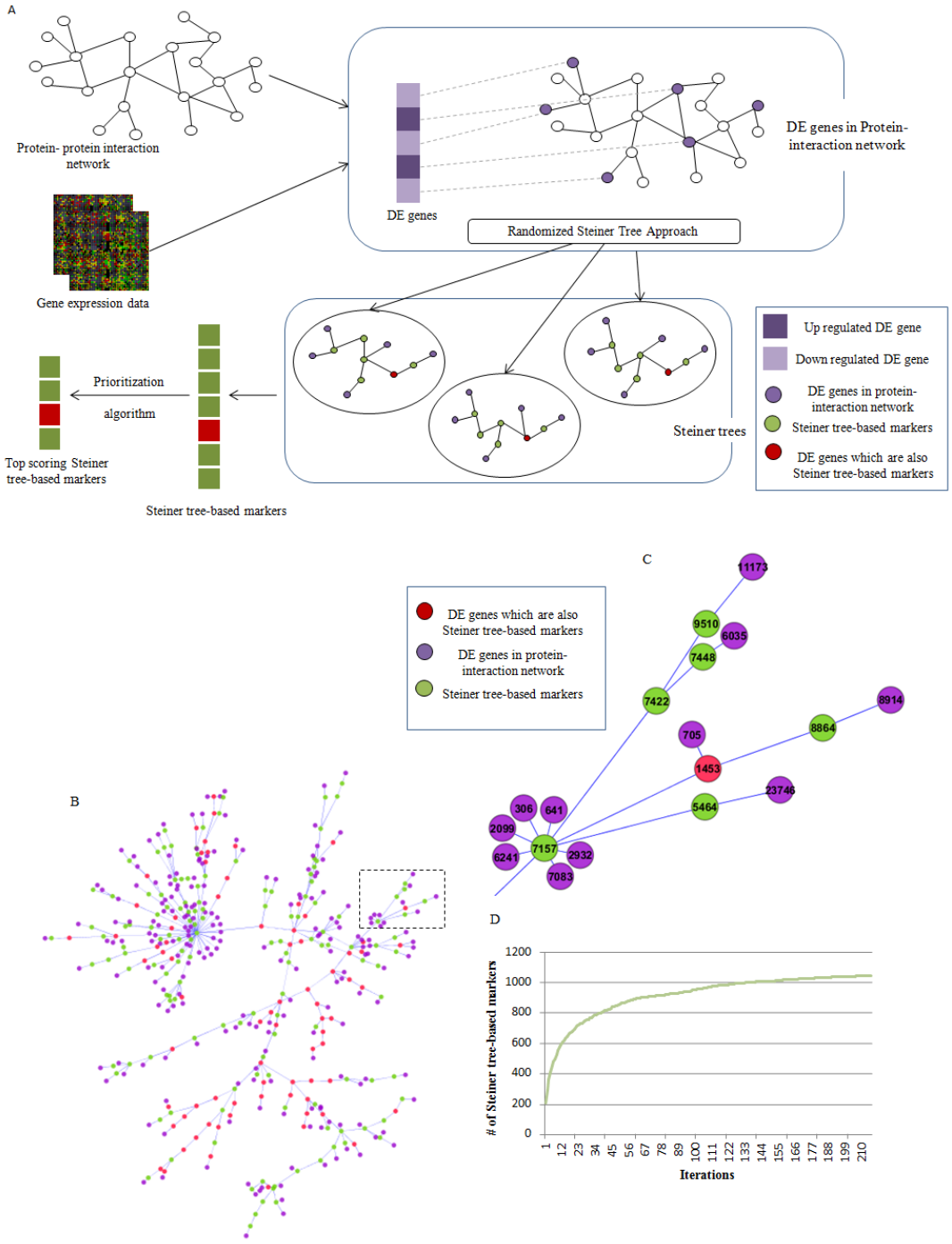


Fig. 1. A) Overview of the randomized Steiner tree-based approach. B) A single Steiner tree for van de Vijver dataset in Chuang-PPI network. C) A larger figure for the highlighted part of the tree. The ids of the vertices are in entrez gene id. D) After generating about 300 Steiner trees the number of STMs saturates to 1047 genes.

2 RESULTS

2.1 Overview of our approach

This study mainly used two breast cancer microarray datasets by Wang *et al.* (2005) and van de Vijver *et al.* (2002), herein referred to as Wang dataset and van de Vijver dataset, respectively, and two PPI networks, PINA-PPI (Wu *et al.*, 2010) and the network used by Chuang *et al.* (2007) (Chuang-PPI). PINA-PPI provides slightly better results in general, but Chuang-PPI is used to compare the

results with Chuang *et al.* (2007). Unless otherwise stated, in this section PPI referred to Chuang-PPI network.

An overview of our approach is shown in Figure 1. Briefly, the method starts with selecting a set of genes that are differentially expressed between metastatic and non-metastatic patients, and then connects them with a Steiner tree in the PPI network. We also use a randomized approach to obtain multiple Steiner trees in order to cover all the possible paths between the DE genes. Finally, genes on the paths are ranked by a prioritization criterion, and those passing a certain cutoff are selected as candidate markers. For example, for van de Vijver dataset, a single Steiner tree uses 136 vertices to connect 333 DE genes (Figure 1B-C). Among the DE genes, 264 are leaf nodes and the other 69 are internal nodes. As these internal DE genes are important in connecting the remaining DE genes, we combine the Steiner vertices with these internal DE genes as potential biomarkers for breast cancer metastasis. The total number of biomarkers saturates to 1047 after generating a sets of Steiner trees (Figure 1D). Finally, these genes are ranked and 225 genes are selected as the top-scoring markers. In subsequent subsections, both the intermediate vertices (Steiner tree-based marker, or STM for short) and the top scoring intermediate vertices (top-scoring Steiner tree-based marker, or t-STM for short) are further analyzed as potential candidates for biomarkers.

2.2 Stability of markers across different datasets

We first examined the stability of (top-scoring) Steiner tree-based markers across different datasets. The numbers of genes selected by our method and their overlap for Wang and van de Vijver datasets using the two PPI networks are listed in Table 1. For DE genes the overlap is 7.8% while the overlap between different datasets with our method varies from 20.4% to 28.7%. Chuang *et al.* (2007) has 12.7% overlap while the gene markers from the two original studies have only 2% overlap (van de Vijver *et al.*, 2002; Wang *et al.*, 2005). In addition, the STM/t-STM markers selected from the same data set using different PPI networks have a similar level of overlap (data not shown). Therefore it is evident that in our approach the stability of potential marker genes across different datasets increases significantly than previous studies.

It is also worth mentioning that the overlap between the t-STMs represents an advance from the STMs. Taking the Chuang-PPI for example, with 410 common STMs between the two datasets, we would only expect $\frac{410\times225\times194}{1047\times1100} = 16$ common genes between the two sets of t-STMs if they were randomly selected from the corresponding STMs, in contrast to the 71 observed (p-value $<$ 3E-24, multivariate hypergeometric distribution, see Methods). Similarly, for PINA-PPI, the two sets of t-STMs share 87 common genes, while 14 are observed (p-value $<$ 1E-45).

2.3 Functional enrichment and pathway analysis of Steiner tree-based marker

To reveal the biological functions of the candidate biomarkers, we used the online bioinformatics resource DAVID to analyze the enriched gene ontology biological processes and KEGG pathways within the STMs and t-STMs (Huang *et al.*, 2008). This analysis showed that genes selected by our method are significantly enriched in functional terms that are well known to be involved in cancer carcinogenic process, including cell cycle, apoptosis, DNA repair, gene expression, MAPK, ErbB and P53 signaling pathways. A list of the enriched terms and their enrichment scores are listed in Figure 2. These enriched functional terms and pathways have strong consistency with several previous studies (Chuang *et al.*, 2007; Hwang *et al.*, 2008; Yao *et al.*, 2010). Figure 2 also shows that the DE genes have much lower enrichment scores compared to STMs and t-STMs in almost all pathways listed. Starting with a set of DE genes with low enrichment scores, our method identified a set of genes with very high enrichment scores in cancer carcinogenic process. The enrichment scores difference for different functional categories between DE genes and STMs/t-STMs are very similar with Figure 2 for PINA-PPI network with these two datasets, signifying the robustness of our method.

2.4 Steiner Tree-based markers correspond to biomarker of cancer

Next, we evaluated our method using two lists of known breast cancer susceptibility genes. First, we used 60 genes collected by Chuang *et al.* (2007) from Online Mendelian Inheritance in Man (OMIM). As shown in Figure 3, DE genes and potential markers from the previous studies have lower percentage of overlaps with the known susceptibility genes than STMs and t-STMs for both Wang and van de Vijver datasets. In addition, the percentages of cancer susceptibility genes are increased in t-STMs than in STMs (p-value =0.002 and 0.06 for Wang and van de Vijver datasets, respectively).

In addition, we collected 288 breast cancer susceptibility genes from Genetic Association Database of Disease from DAVID. Among the STMs, 6.9% and 8.3% are known disease genes for

Table 1. The percentage of overlap and p-values between van de Vijver and Wang datasets in different methods

	van de Vijver dataset	Wang dataset	Number of common genes (% of overlap)	p-value
Chuang-PPI (STM)	1047	1100	410(23.6%)	6.2E-159
PINA-PPI (STM)	932	1135	370(21.8%)	7.7E-138
Chuang-PPI (t-STM)	225	194	71(20.4%)	5.0E-072
PINA-PPI (t-STM)	205	185	87(28.7%)	9.1E-106
DE genes	333	319	47(7.8%)	1.3E-019
Result in (Chuang *et al.*, 2007)	618	906	175(12.7%)	5.6E-054
Result in (van de Vijver *et al.*, 2002; Wang *et al.*, 2005)	70	76	3(2.1%)	.03

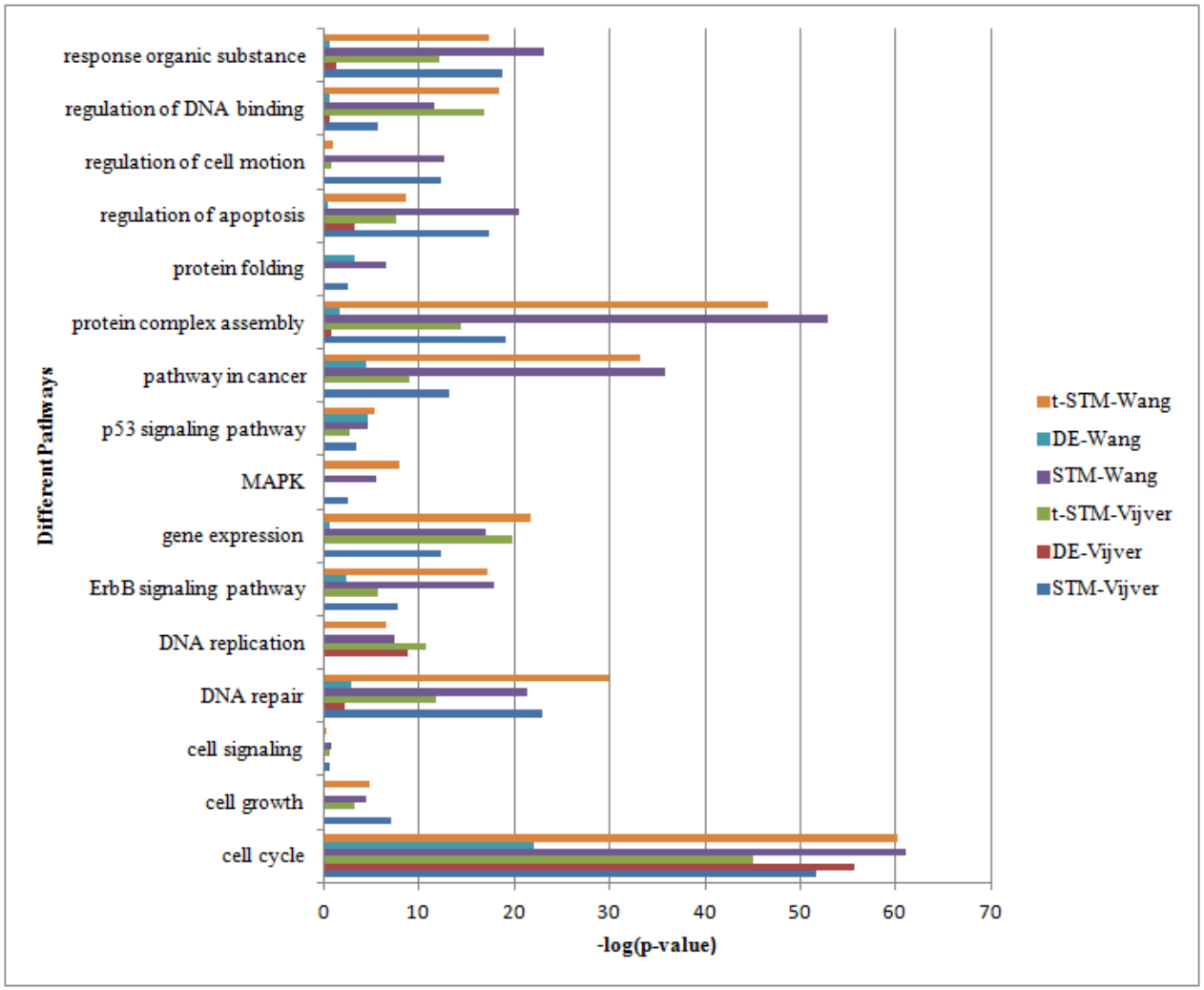


Fig. 2. Enriched biological processes and pathways of STMs, t-STMs and DE genes for van de Vijver and Wang dataset with Chuang-PPI network. Statistical significance (p-value) was estimated using the cumulative hyper-geometric distribution test (Fisher's exact test), using as background the largest connected component of the PPI for the STMs and t-STMs, and all genes on the chip for the DEs.

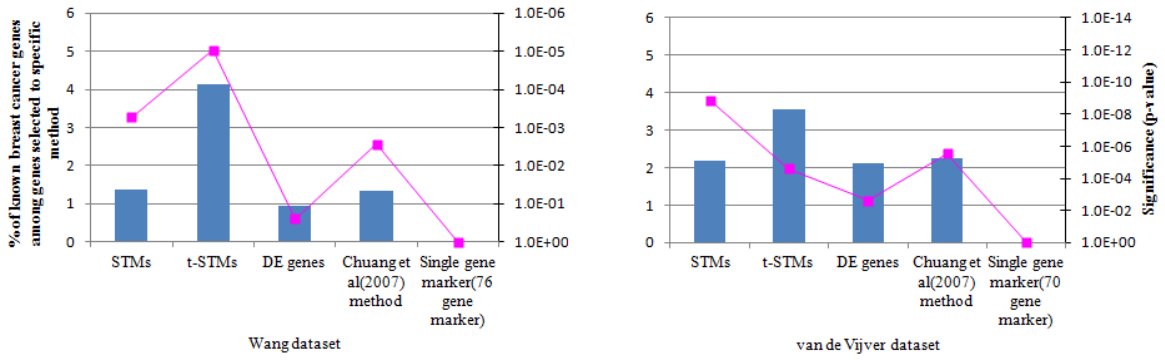


Fig. 3. Overlap between STMs, t-STMs, DE genes, Chuang *et al.* (2007) method and corresponding original studies and the selected 60 breast cancer susceptibility genes. Bars represent percentages and lines depict p-values.

Wang and van de Vijver datasets, respectively. Concentrations of known disease genes in the t-STMs are 15.6% and 11.3%, for the two datasets respectively, significantly higher than in the STMs (p-value = 7E-7 and 0.03, respectively). For both datasets, genes selected by our method clearly outperformed DE genes(5.0% and 5.9% for Wang and van de Vijver datasets respectively). Evaluation results using PINA-PPI show similar results.

2.5 Steiner tree-based markers improves the cross-dataset classification accuracy

We tested whether the STMs can be used to improve the prediction accuracy of breast cancer metastasis. To this end we use the STMs or t-STMs as features to train Bayesian logistic regression and support vector machine (SVM) classifiers (Genkin *et al.*, 2007; Hall *et al.*, 2009) to separate metastatic from non-metastatic patients, using the implementation in WEKA (version 3.6.6) with default parameters (Hall *et al.*, 2009). For comparison, we also constructed classifiers using only DE genes, and using the combination of DE and STM/t-STM genes. To evaluate the power of the candidate markers in predicting metastasis in an unbiased fashion, we focused on cross-dataset tests, where the genes selected from one dataset were used to construct classifiers for the other dataset. Classification performance was estimated 100 times using 5-fold cross-validation where iteratively one-fifth of the data were used for testing and four-fifth were for training. Figure 4 shows the classification accuracy, measured by AUC (area under ROC curve), achieved by different gene sets. As can be seen, STMs resulted in better accuracy than DEs in all cases. For example, on van de Vijver dataset, STM-based Bayesian logistic regression classifier performed better than DE-based classifiers in 96 of the 100 runs (p-value = 1E-31, paired t-test), and STM-based SVM classifier outperformed DE-based SVM in 87/100 of runs (p-value = 6E-20). For Wang dataset, the corresponding numbers are 99/100 (p-value = 3E-37) and 100/100 (p-value = 4E-43). On the other hand, combinations of DEs and STMs do not seem to improve accuracy significantly, suggesting that the pathway information contained in the DE genes are redundantly present in the STMs. The t-STMs, on the other hand, usually result in an accuracy similar to or slightly lower than that of DEs, which may be attributed to the relatively small sizes of t-STMs. Combinations of DEs and t-STMs do bring the accuracy to usually above the level of DE-based classifiers, and in one case significantly increased the accuracy (van de Vijer dataset, Bayesian logistic regression), indicating that t-STMs can provide complementary and orthogonal information to DEs.

To further understand the distinction between DEs and STMs, we analyzed the regression coefficients (weights) of the DE and STM genes in the corresponding Bayesian logistic regression classifiers. We sorted the genes by the magnitude of their coefficients in the classifier and selected the top ten genes with the largest absolute values. Table 2 shows these genes and their weights. To confirm that the STMs do represent biomarkers for metastasis, we searched PubMed abstracts with the combination of the gene name and "metastatic" or "metastasis". Table 2 shows the number of publications retrieved as a result of the search. Surprisingly, only about half of DE genes were associated with metastasis in PubMed, and most of them were supported by a few publications. In contrast, 90% of the STMs are associated with metastasis and many of them are well studied, including Brms1, which encodes the breast cancer metastasis-suppressor 1 protein (Shevde *et al.*, 2002), and Mmp1/Mmp9, which belong to the matrix metalloproteinase (MMP) family and are well-known for their role in metastasis (Kleiner and Stetler-Stevenson, 1999).

Interestingly, three of the top-ten genes in Wang-STM classifier, Bnip3, Phgdh, and Aars, are also top-ten genes in Wang-DE classifier. As the genes used by the Wang-STM classifier were actually obtained from the van de Vijver dataset (for cross-data validation), it seems that the STM genes are also robust on the classifier level. In particular, the Aars gene has the highest weight in both Wang-DE and Wang-STM classifiers, but is not known to be involved in metastasis by PubMed search. Therefore, it may represent an interesting biomarkers for further investigation.

2.6 Validate the method with another dataset

Finally, we validated our method with the dataset from Desmedt *et al.* (2007). Starting with 354 DE genes, we obtained 1263 (1291) STMs and 157 (147) t-STMs using the Chuang (PINA) PPI. Similar to the results for the previous two datasets, the genes selected from this dataset are also significantly enriched in many functional terms and pathways that are well known to be associated with breast cancer, and cover significant numbers of known breast cancer susceptibility genes. Genes selected by our method for this dataset have a significant overlap with the other two datasets mentioned. As shown in Tables 3 and 4, STMs and t-STMs have between 24.6% and 31.2% overlap, while the DEs have a significantly lower (maximum 13.9%) overlap, indicating that the results from our method are general to different datasets.

3 METHODS

3.1 Expression data and protein-protein interaction networks

Three breast cancer microarray datasets were used in this study (van de Vijver *et al.*, 2002; Wang *et al.*, 2005; Desmedt *et al.*, 2007). The primary breast tumor samples in three datasets were obtained from 295, 286 and 198 patients, respectively; among them, 78, 106 and 35 patients had metastasis during follow up visit within 5 years of checkup. The microarray platform used in the last two datasets (Desmedt *et al.*, 2007; Wang *et al.*, 2005) was Affymetrix HG-U133a, and the first dataset (van de Vijver *et al.*, 2002) was Agilent Hu25K. Two human protein-protein interaction networks were used. The first network, PINA, contains 10,920 proteins and 61,746 binary connections (Wu *et al.*, 2010). The other network was compiled by Chuang *et al.* (2007), which integrated 57,235 interactions among 11,203 proteins from different sources. For our approach we only considered the largest connected component in the PPI networks which contain 10,794 and 10,770 proteins for the Chuang and PINA PPI networks respectively.

3.2 Selecting differentially expressed (DE) genes

Significant Analysis of Microarray (SAM) (Tusher *et al.*, 2001) was used to select DE genes. Patients were divided into two phenotypes (metastatic vs. non-metastatic) based on the follow-up visit within 5 years of checkup. To select DE genes, FDR 20%, 8.21% and 0.65% were used as cutoffs for Desmedt, Wang and van de Vijver datasets, respectively (Desmedt *et al.*, 2007; Wang *et al.*, 2005;

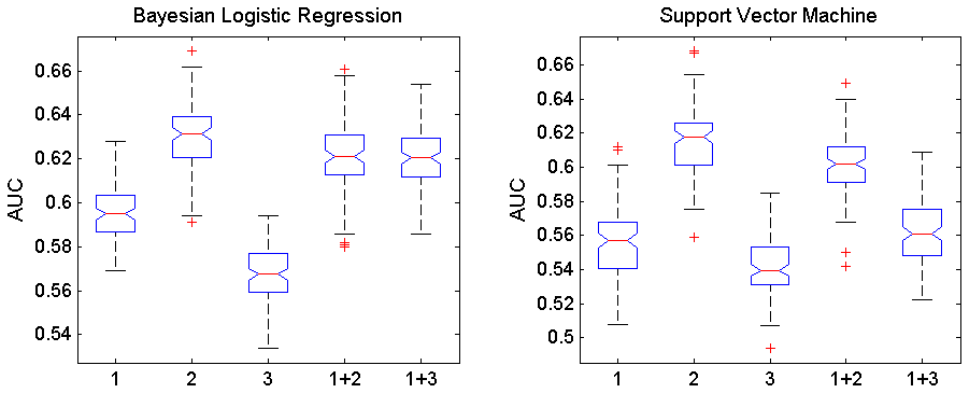
Table 2. Literature validation of top marker genes chosen by Bayesian logistic regression.

Wang-DE			Wang-STM			van de Vijver-DE			van de Vijver-STM		
Gene	Weight	#Papers	Gene	Weight	#Papers	Gene	Weight	#Papers	Gene	Weight	#Papers
Bnip3	3.74	17	Ccl2	-4.65	88	Adam8	2.62	7	Mmp9	3.63	231
Eif4ebp1	2.33	2	Mmp1	4.18	64	Cse1l	-3.14	7	Brms1	3.44	75
Dtl	2.21	1	Myb	-4.15	45	Ptk6	2.79	4	Igfbp5	4.26	10
Mad2l1	-2.52	1	Gstt1	-3.74	20	Cdc27	-2.15	2	Cse1l	-2.64	7
Phgdh	-3.98	1	Bnip3	4.51	17	Bub1b	2.67	1	Cd47	-3.14	6
Racgap1	2.38	1	Sell	3.89	9	Acbd3	2.14	0	Itga2	-3.94	5
Aars	-4.65	0	Asns	3.84	2	Ctdsp1	2.67	0	Sgk1	-2.70	4
Aldh3a2	-3.24	0	Shc1	3.63	2	Mtfr1	-2.41	0	Degs1	2.80	1
Hmgb3	2.44	0	Phgdh	-3.41	1	Mtm1	-2.29	0	Pdhb	-2.83	1
Tpsb2	2.72	0	Aars	-5.84	0	Zcchc8	-4.09	0	Ldoc1	-3.01	0

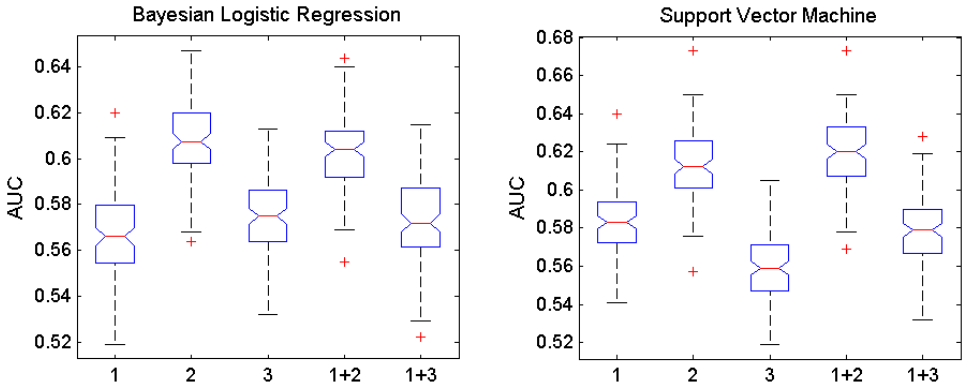
van de Vijver *et al.*, 2002), and a total of 319, 354 and 333 genes were selected respectively. Different FDR cutoffs were chosen for different datasets in order to obtain comparable numbers of DE genes.

3.3 Randomized Steiner tree approach

The main idea in our approach is to find a small number of genes that can connect all the DE genes into a singly connected component in a PPI network. (As a PPI network may contain several disconnected components, we only considered the DE genes that fall into the largest connected component of the PPI network.) Intuitively, we



(a) Classification results on Wang dataset



(b) Classification results on van de Vijver dataset

Fig. 4. Classification accuracy of DE (1), STM (2), t-STM (3) genes and their combinations based on AUC metric with Bayesian logistic regression and Support Vector Machine are shown. In all cases, cross-data validation was performed, meaning that the features for constructing the classifiers of one dataset were taken from the other dataset.

Table 3. The percentage of overlap and p-values between Desmedt and Wang dataset

	Desmedt dataset	Wang dataset	Number of common genes (% of overlap)	p-value
Chuang-PPI (STM)	1263	1100	467(24.6%)	5.3E-180
PINA-PPI (STM)	1291	1135	517(27.1%)	1.8E-210
Chuang-PPI (t-STM)	157	194	76(27.6%)	3.0E-096
PINA-PPI (t-STM)	147	185	79(31.2%)	1.8E-106
DE genes	354	319	77(12.9%)	2.0E-040

Table 4. The percentage of overlap and p-values between Desmedt and van de Vijver dataset

	van de Vijver dataset	Desmedt dataset	Number of common genes (% of overlap)	p-value
Chuang-PPI (STM)	1047	1263	482(26.4%)	5.2E-199
PINA-PPI (STM)	932	1291	468(26.7%)	8.0E-208
Chuang-PPI (t-STM)	225	157	86(29.0%)	5.8E-108
PINA-PPI (t-STM)	205	147	79(28.9%)	1.5E-101
DE genes	333	354	84(13.9%)	2.7E-052

are looking for the most parsimonious solution, that is, a spanning tree that connects the DE genes with the fewest additional genes. In graph theory, this maps to the well-known Steiner tree problem. Formally, the Steiner tree for an edge-weighted graph G= (V, E, w) and a subset of vertices R⊆V is a minimum-weight connected tree T, with vertices U⊆V and edges S⊆E that spans all vertices in R. Here the vertices in R are known as terminal vertices and U-R as Steiner vertices. For an unweighted graph G, the problem then becomes to find the minimum number of vertices that can connect all the vertices in R through a tree in G. The Steiner tree problem is NP-hard (Vo, 1992). We implemented a polynomial-time 2-approximation shortest path heuristic algorithm of Steiner tree problem (Rayward-Smith, 1983).

After obtaining a Steiner tree, we considered both the Steiner vertices and the DE genes that are internal nodes in the tree as potential biomarkers (Steiner tree-based markers, or STMs for short), as they play important roles in forming connections among multiple DE genes. Next, we designed a randomized algorithm to obtain multiple Steiner trees, for two reasons. First, as the heuristic algorithm does not guarantee optimality, by obtaining multiple solutions we increase the chance of finding the optimal Steiner vertices. Second, multiple solutions with similar qualities may represent alternative or redundant pathways that cannot be covered by a single Steiner tree. To this end, we assign to each edge of the PPI network a random weight between 0.99 to 1, and run the Steiner tree algorithm. These random weights effectively break ties, so that if there are two paths with the same weight in the original network, one path will be chosen randomly. This procedure was repeated multiple times with different random weights, until the total number of unique STMs converges approximately. Depending on the PPI network and microarray data, the rate of new coming STMs reduced significantly after 200-300 iterations (for example, see Fig 1D).

3.4 Top-scoring Steiner tree-based markers

To further prioritize the STMs, we scored them using the following method. As disease genes play critical role in connecting DE genes in the PPI network, we calculated for each STM its average distance to its nearest n (n = 5 in our experiment) DE genes. Then the score for each STM is defined as the reciprocal of that average distance and a certain number of top-scoring Steiner tree-based markers (t-STMs) were identified using a user-defined cutoff.

3.5 Statistical test

Let N be the number of genes in the PPI network, m and n the sizes of two gene sets, o the size of the overlap, the percent overlap between the two gene sets is calculated as $100 \times o/(m+n-o)$, and the statistical significance (p-value) of the overlap is calculated using the cumulative hyper-geometric distribution:

$$p = 1 - \sum_{i=0}^{o-1} C(m,i)C(N-m,n-i)/C(N,n), \qquad (1)$$

where *C(n, k)* is the binomial coefficient.

To calculate the statistical significance of the overlap between two sets of t-STMs using the corresponding STMs as background, we use the multivariate hyper-geometric distribution. Let the sizes of two STM sets be X and Y, respectively, and the size of the overlap between them be O. Let the sizes of the two corresponding t-STM sets be x and y, respectively and the size of the overlap between them o. The probability that this level of overlap in t-STMs can be achieved by randomly drawing x and y genes from the two corresponding STMs can be calculated

$$p = 1 - \sum_{k=o}^{min(x,y,O)} \sum_{i=k}^{min(x,O)} \sum_{j=k}^{min(y,O)} H(i, X, x, O) M(o, j-o, y-j, i, O-i, Y-O) \quad (2)$$

by Equation (2), where H is the probability distribution function for hyper-geometric distribution and M is that for multivariate hypergeometric distribution.

4 CONCLUSION

In this article we proposed a randomized Steiner tree-based approach that integrates a PPI network and gene expression microarray data for biomarker discovery in breast cancer metastasis. The genes selected by our method are significantly enriched in functional categories and pathways that are known for cancer development. Furthermore, a significant portion of selected genes by our method are already known for breast cancer susceptibility. We applied the method to three different breast cancer microarray data and two different PPI networks. For all combinations of microarray and PPI datasets our approach has similarly significant results. The reproducibility across different datasets also increases significantly in both genomic and pathway level compared to previous studies. Finally, Steiner tree-based markers, significantly increase cross-dataset classification accuracy. Thus the method proposed in this article validates the hypothesis that disease causal genes play a role in connecting differentially expressed genes, and opens a new possibility to identify the inner dynamics and biomarker of breast cancer progression.

ACKNOWLEDGMENT

Funding: SC3GM086305, U54CA113001, R01CA152063

REFERENCES

Chuang, H., Lee, E., Liu, Y., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, **3,140**.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G., Foekens, J. A., Cardoso, F., Piccart, M. J., Buyse, M., Sotiriou, C., and TRANSBIG Consortium (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **13**(11), 3207–3214.

Dezso, Z., Nikolsky, Y., Nikolskaya, T., Miller, J., Cherba, D., Webb, C., and Bugrim, A. (2009). Identifying disease-specific genes based on their topological significance in protein networks. *BMC systems biology*, **3:36**.

Genkin, Alexander, Lewis, David, D., Madigan, and David (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, **49**(3), 291–304.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, **286**(5439), 531–537.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protocols*, **4(1):44-57**.

Hwang, T., Tian, Z., Kuang, R., and Kocher, J.-P. (2008). Learning on weighted hypergraphs to integrate protein interactions and gene expressions for cancer outcome prediction. In *ICDM '08. Eighth IEEE International Conference on Data Mining*, pages 293 –302.

Kleiner, D. E. and Stetler-Stevenson, W. G. (1999). Matrix metalloproteinases and metastasis. *Cancer Chemotherapy and Pharmacology*, **43**, S42–S51.

Ma, S., Shi, M., Li, Y., Yi, D., and Shia, B.-C. (2010). Incorporating gene co-expression network in identification of cancer prognosis markers. *BMC Bioinformatics*, **11**(1), 271.

Pavlidis, P., Qin, J., Arango, V., Mann, J. J., and Sibille, E. (2004). Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex. *Neurochemical Research*, **29**(6), 1213–1222.

Rayward-Smith, V. J. (1983). The computation of nearly minimal Steiner trees in graphs. *Internat J. Math. Ed. Sci. Tech*, **14**, 15–23.

Shevde, L. A., Samant, R. S., Goldberg, S. F., Sikaneta, T., Alessandrini, A., Donahue, H. J., Mauger, D. T., and Welch, D. R. (2002). Suppression of human melanoma metastasis by the metastasis suppressor gene, brms1. *Experimental Cell Research*, **273**(2), 229 – 239.

Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer. *The New England journal of medicine*, **360**(8), 790–800.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, **27**(2), 199–204.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(38), 13544–13549.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98**(9), 5116–21.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van Der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England journal of medicine*, **347**(25), 1999–2009.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.

Vo, S. (1992). Steiner's problem in graphs: heuristic methods. *Discrete Applied Mathematics*, **40**(1), 45 – 72.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**(9460), 671–679.

Wei, Z. and Li, H. (2007). A Markov Random Field Model for Network-based Analysis of Genomic Data. *Bioinformatics*.

Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, **11**(5).

Yao, C., Li, H., Zhou, C., Zhang, L., Zou, J., and Guo, Z. (2010). Multi-level reproducibility of signature hubs in human interactome for breast cancer metastasis. *BMC systems biology*, **4:151**.