A COMPARATIVE STUDY OF BREAST CANCER MICROARRAY GENE EXPRESSION PROFILES USING LABEL PROPAGATION

TaeHyun Hwang and Rui Kuang

Department of Computer Science and Engineering, University of Minnesota
200 Union Street SE, Minneapolis, MN 55455
thwang@cs.umn.edu, kuang@cs.umn.edu

ABSTRACT

A challenge in using microarray gene expression profiles to study breast cancer is to analyze the inconsistent discoveries made from independent microarray datasets. The inconsistency is often related to the tuning of those sophisticated strategies needed for taking into account the dependence among the genes in the analysis as well as the difference between the platforms and the protocols used for generating the datasets. In this paper, we use a simple graph labeling algorithm which can capture the dependency among the genes to study breast cancer microarry data. We perform a comparative study of breast cancer metastasis on two datasets using the graph labeling algorithm and the standard statistics of correlation coefficients. We show that our algorithm predicts more consistent marker genes and pathways enriched by the marker genes on the two datasets than the correlation-coefficient statistics.

1. INTRODUCTION

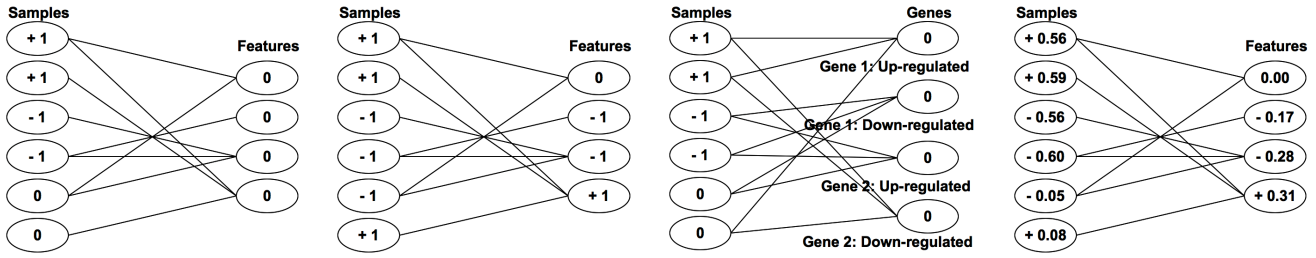
Recently, many microarray gene expression profiles have been made available for investigating the genetic associations of various human diseases. Identification of genetic markers of the diseases can provide useful information for both the treatment and the etiology. It has been shown that the discovered genetic markers of disease can potentially provide better prognosis and diagnosis than the currently available clinical measures for risk assessment in patients of various diseases [8, 4, 7, 5, 2]. A number of disease markers have been discovered through genome-wide expression profiles using various computational methods such as Bayesian networks, support vector machines and other statistical methods [8, 4, 7, 5]. In the study of breast cancer, [4] and [7] independently identified two sets of marker genes related to the metastasis of breast cancer using large scale gene expression profiles produced in two different microarray experiments. However, there are only three genes in common between the two sets of the marker genes. Moreover, those two lists of marker genes do not include any genes with known breast cancer mutations such as P53, KRAS, HRAS, HER-2/neu and PIK3CA [4, 7, 2]. One possible explanation is that the two sets of gene expression profiles use different microarray platforms, and the difference introduced by the experiment techniques might affect expression patterns used for the selection of marker genes. Another possibility is when different training sets of patients and different methods are used, achieving a stable selection of marker genes becomes a major challenge due to the intensive computational need for searching through all possible combinations of the genes. In reality, the lists of marker genes discovered from different gene expression profiles on breast cancer are rarely overlapped, although the genes are often involved in common pathways[2, 3].

In this paper, we apply a simple graph learning algorithm, Network Propagation, which can capture the dependance among the genes, to find a stable global optimal solution. We show that our algorithm can find more consistent sets of marker genes which share in common several important biological networks and pathways associated with breast cancer using two heterogeneous breast cancer datasets.

2. METHODS

In this study, we formulate a graph labeling problem for associating marker genes with specific phenotypes in a disease context [6]. We use labeled and unlabeled samples to classify the gene features into positive, negative or neutral classes (Figure 1A and 1B). The positively classified gene features and negatively classified features are candidate marker genes. In the bipartite graph, *feature vertices* represent up/down-regulated genes; *object vertices* represent labeled and unlabeled samples, connected to the feature vertices by weighted edges. The object vertices are labeled with -1/+1 if the label is known, 0 otherwise. Every gene is represented by two vertices, up-regulated and down-regulated; each sample will be connected to either the up-regulated vertex or the down-regulated vertex with an edge weighted by the expression level (Figure 1C). The global optimal solution of this problem can be found by a network propagation algorithm (Figure 1D). This learning algorithm can capture the interactions between both samples and gene features by exploring the global structure of the bipartite graph, based on a "cluster assumption": those samples in the same class tend to be heavily connected to a common set of features; those features that can characterize a class tend to be heavily connected to the samples in the class. The semi-supervised learning problem is about how to learn the best labels on all the vertices, given the bipartite graph and the known labels.

The network propagation utilizes phenotype labels to achieve a global optimum for classifying "positive", "negative" and "neutral" features along with test samples in one semi-supervised learning procedure.The nature of our graph-based learning algorithm captures dependency between all genes simultaneously by exploring the graph structure, which is essentially a non-linear method for selecting genes. The network propagation propagates the label information of every genes to its neighbors in the other gene sets. This propagation process will leverage the activation values of the genes in a densely connected neighborhood.



(A) Initial labels (B) Optimal labels (C) Graph with Gene Expressions (D) Predicted labels

Figure 1: **Feature classification on a bipartite graph.** This example shows a toy graph with 6 sample vertices and 4 feature vertices. All the edges are assumed uniformly weighted. (A) Four samples are initially labeled according to their phenotype class; the other two and all the feature vertices are unknown. (B) The two feature vertices strongly connected to the negative vertices are labeled negative, the one feature vertex strongly connected to the positive vertices is labeled positive, and the one that is connected to both classes is assigned 0. The two unlabeled samples are also labeled according to their connections in the graph. (C) The prediction scores (activation values) produced by Network Propagation with $\alpha = 0.5$ and 1000 iterations. All the nodes are correctly labeled; note that the labels are relaxed into real numbers. (D) A bipartite graph with vertices of gene expressions; the edge weights are the absolute expression levels of the genes.

3. EXPERIMENTS

We use the network propagation algorithm and correlation coefficients to analyze two expression profiles of lymph-node-negative primary breast cancer patients used in previous studies [4, 7]. For our convenience, we use "Rosetta" and "Wang" to refer to the two datasets respectively. The two datasets are generated from primary breast tumor tissues of different patient groups on different mircroarry platforms.

3.1 Data preparation

The Rosetta dataset hybridized to Agilent oligonucleotide Hu25K microarrays. The microarray gene expression profile measures the expression levels of 24,481 genes from 97 patients. We analyze all 24,481 gene expressions in our experiments. The details for quantization and normalization of scanned microarray images are described in [4]. The Wang dataset hybridized to the Affymetrix oligonucleotide microarray U133a GeneChip [7]. The expression of 22,283 transcripts are collected from total RNA of frozen samples from 286 lymph-node-negative breast cancer patients. We treat each probeset as a separate gene and normalize gene expression values using median value for each gene. Then we rescale them by log_2 ratio. In this analysis, we use all patients label information for identifying marker genes. Note to remove any other artificial effects in our experiments, all the genes in the two dataset are used as candidates of marker genes without any pre-pruning. Thus, it is very difficult to get overlapped predictions on the two datasets, given the number of genes in the two datasets.

3.2 Identification of marker genes and over-represented pathways

To identify gene markers, we use the network propagation algorithm and the correlation-coefficient statistics. After running the network propagation, we list the top-ranked genes by the absolute values of Z-scores calculated from the activation values. For more details, please refer [6]. With the standard statistic method, we calculate correlation coefficients using the label information. We then list top-ranked genes based on the absolute value of correlation coefficients. Note that we use top-200 ranked genes for both methods on the Rosetta and Wang's datasets. To identify over-representation of biological pathways, we use Ingenuity software (version 5.5) with the sets of top-200 genes identified by the Network Propagation and the standard statistic measurement using correlation coefficients from the Rosetta and Wang's datasets. The software automatically preprocess the gene list to select genes eligible for pathway analysis. In our gene lists, only 91 (Rosetta) and 144 (Wang) out of 200 selected genes using the Network Propagation and 85 (Rosetta) and 115 (Wang) out of 200 selected genes using the standard statistic measurement are selected by the software and then used as input to search for the biological networks provided by the software. The biological networks identified by the software are assessed in the context of general functional classes in Gene Ontology. We only investigate the functions involving at least two selected genes for both the Rosetta dataset and the Wang dataset and scoring a p-value less than 0.01. If any of these two criteria are not satisfied, we remove the function in our list. We also examine the enriched biological networks using the selected gene markers identified by the software. The software uses our selected genes as "seed" to find relevant networks based on the knowledge base in the system and then return the enriched biological networks. Each network contains a set of genes which share the same function and interact with each other.

3.3 Results of Network Propagation and the correlation-coefficient statistics

We compare the marker genes identified by Network Propagation and correlation-coefficient statistics from the Rosetta and Wang's datasets. We also analyze the enriched biological networks and the over-represented pathways identified by the software on the sets of marker genes.

3.3.1 Compare the marker genes

In the list of marker genes identified by Network Propagation, many genes are related to tumor aggression and metastasis, for example baculoviral IAP repeat-containing 5 (BIRC5) and matrix metallopeptidase 9 (MMP9) are both included [2]. Network Propagation also detects one well known breast cancer susceptibility gene, ER-α (ESR1) [4], in the top-200 genes on the Rosetta dataset. Network Propagation detects another well known breast cancer susceptibility genes in the top-200 genes on Wang's dataset, the phoshpoinositide-3-kinase catalytic subunit (PIK3CA). Some mutations in PIK3CA are associated with constitutive

Table 1: **Comparison of the enriched pathways identified by Network Propagation (NP) and the Correlation Coefficients (CC).** The 17 over-represented functions by the marker genes are analyzed. The significance value associated with each function is measured by p-value calculated using the right-tailed Fisher Exact Test. We only list the functions which have p-value less than 0.01 ($p < 0.01$) measured as the enrichment by the top-200 genes using the two methods on Rosetta and Wang datasets. A cross "X" denotes that the funcntion is identified from the dataset the with corresponding method. Ten out of the 17 functions are in common for all the cases. Five additional functions reported by Network Propagation and two additional functions are reported by correlation-coefficient statistics.

Molecular and Cellular Function	Rosetta		Wang		Description
	NP	CC	NP	CC	
Cellular Assembly and Organization	X	X	X	X	Subcellular components that are involved in cellular organization and assembly of cellular substructures.
DNA Replication, Recombination, and Repair	X	X	X	X	The replication, recombination and repair of DNA.
Cell Cycle	X	X	X	X	Functions and stages of the cell cycle including cell division
Cell Death	X	X	X	X	Cellular death and survival.
Cellular Movement	X	X	X	X	Functions associated with movement and localization of cells.
Molecular Transport	X	X	X	X	The intra- and extracellular movement of molecules, including small molecules, ions, DNA, RNA, protein, lipids and carbohydrates.
Cellular Development	X	X	X	X	The development and differentiation of cells.
Cellular Growth and Proliferation	X	X	X	X	The growth and proliferation of cells.
Cell-to-Cell Signaling and Interation	X	X	X	X	Functions that are involved in intercellular interactions such as binding, detachment, communication, pheromone response, and stimulation.
Cell Morphology	X	X	X	X	The morphology of cells.
Small Molecule Biochemistry	X		X		Functions associated with small molecules
Post-Translational Modification	X		X		The modification of proteins after translation
Lipid Metabolism	X		X		Functions associated with the metabolism of lipids
Small Molecule Biochemistry	X		X		Functions associated with small molecules
Cellular Function and Maintenance	X		X		The normal cellular functions that maintain cellular homeostasis.
Cell Signaling		X		X	Functions that are involved in intracellular signaling pathways.
Cellular Compromise		X		X	The damage or degeneration of cells or any process that might compromise the function of the cell.

up-regulation of kinase activity in around 30% of breast cancers [2]. Six genes, BIRC5, FGB, FGG, NMU, VGLL1 and PCSK1, are overlapped between the twos lists of gene markers from the two datasets. BIRC5, which plays a role associated with P53 signaling, is another well known breast cancer susceptibility gene. FGB and FGG are members of FG and react with Fibrin involved with cell death, one of the well known functions associated with cancer disease. When we compare top-300, 500 and 1000 gene, around 10% of genes are overlapped. Note that in the previous work in [7, 4], only three genes out of 70 genes are overlapped. In the list of marker genes identified by Correlation Coefficients, we detect one well known breast cancer susceptibility gene in the top-200 genes on Wang's dataset, v-Ki0ras2 Kirsetn rat sarcoma viral oncogene homolog (KRAS) [4]. No important genes associated breast cancer in the list of gene markers from Rosetta are detected. There is only one gene overlapped between the two lists from the two datasets. We also examine the enriched set of genes retrieved by the software. We compare the enriched sets of genes with the 60 breast cancer susceptibility genes reported by [2]. In the enriched sets of genes identified by Network Propagation, we detect 26 and 32 out of the 60 breast cancer susceptibility genes on the Rosetta and Wang datasets respectively. The correlation-coefficient statistics detect 20 and 34 genes from the Rosetta and Wang datasets that are overlapped with 60 breast cancer susceptibility genes.

3.3.2 Compare the enriched pathways

The functions of the mark genes and the biological networks involving the genes are analyzed. Molecular and Cellular functions are assigned to each marker gene. In Figure 2, we plot the lists of identified functions. Since we have different numbers of genes identified by the two methods, the functions are plotted based on the significance calculated by p-values. Note that we plot the functions which involves at least two genes on both Rosetta and Wang datasets and scoring a p-value less than 0.01. The enriched functions obtained by both methods show strong consistency with those identified by [1, 7], indicating that these processes are significantly involved with the progression of cancer. Although we only have a small number of overlapped genes in the two lists, almost all functions associated with marker genes are matched from the two datasets. Among the 15 functions, nine functions such as cellular growth and proliferation, cell death, cell cycle and etc, are exactly or closely matched with the 21 functions discovered previously in [7]. Among the 12 functions by the standard statistic method, nine functions are overlapped with previous work [7]. Although, the same number of functions are overlapped with previous work [7], Network Propagation produces more a consistent set of enriched functions on the Rosetta and Wang's datasets than the correlation-coefficient statistics. This might indicate that Network Propagation captures the dependence between the genes of similar roles in those biological networks, and thus finds more common pathways associated disease from heterogenous datasets. In Table1, we show the full list of the common functions identified by the two methods on the two datasets.

4. DISCUSSION

In this paper, we apply the network propagation algorithm and the standard statistic method using correlation coefficients to identify marker genes associated with breast cancer metastasis. We compare the marker genes and the biological pathways identified by each method from two heterogenous breast cancer datasets. Our results suggest that one advantage of Network Propagation is that it can identify more consistent marker genes from the heterogenous breast cancer datasets with more enriched biological pathways in common. One possible explanation is that the network propagation algorithm is a simple strategy to consider the dependence between genes and it can discover the set of marker genes with similar functions or closely interacting with some other important genes. In our future work, we plan to design new algorithms which can incorporate prior knowledge from known functions or pathways for more accurate marker gene discovery.

REFERENCES

[1] Hanahan D and Weinberg RA. The hallmarks of cancer. *Cell*, 100:57–70, Jan 2000.

[2] Han-Yu Chuang et al. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3:140, 2007.

[3] Jack X Yu et al. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*, 7:182, 2007.

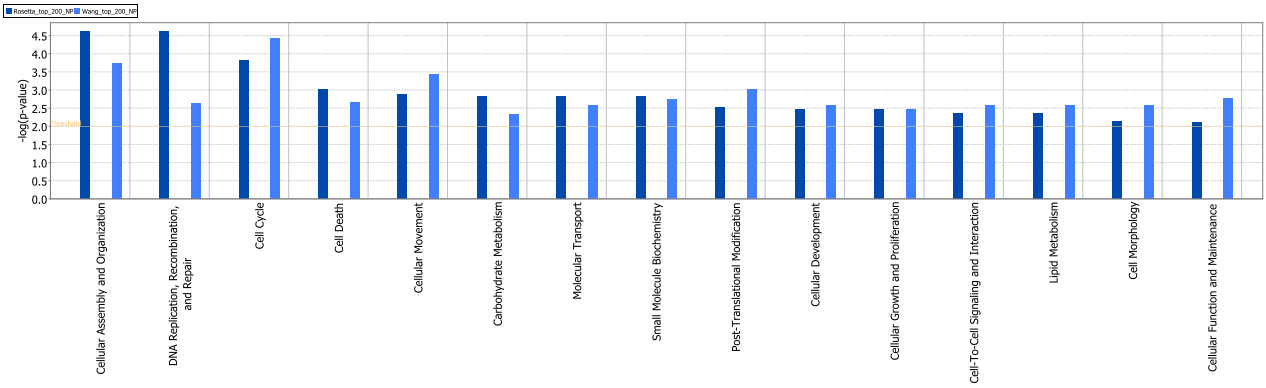
[4] Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[5] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

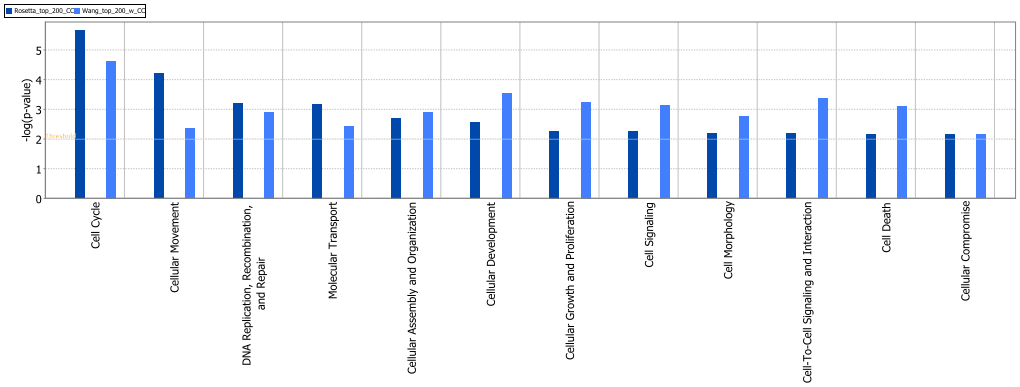
[6] TaeHyun Hwang, Hugues Sicotte, Dennis Wigle, Jean-Pierre Kocher, Vipin Kumar, and Rui Kuang. Identifying clinical and genetic markers of human disease by classifying features on graphs. Technical Report UMN-CS-07-021, University of Minnesota, Twin Cities, 2007.

[7] Y. Zhang A. Sieuwerts M. Look F. Yang D. Talantov M. Timmermans M. Meijer-van Gelder J. Yu Y. Wang, J. Klijn. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365:671–679, 2005.

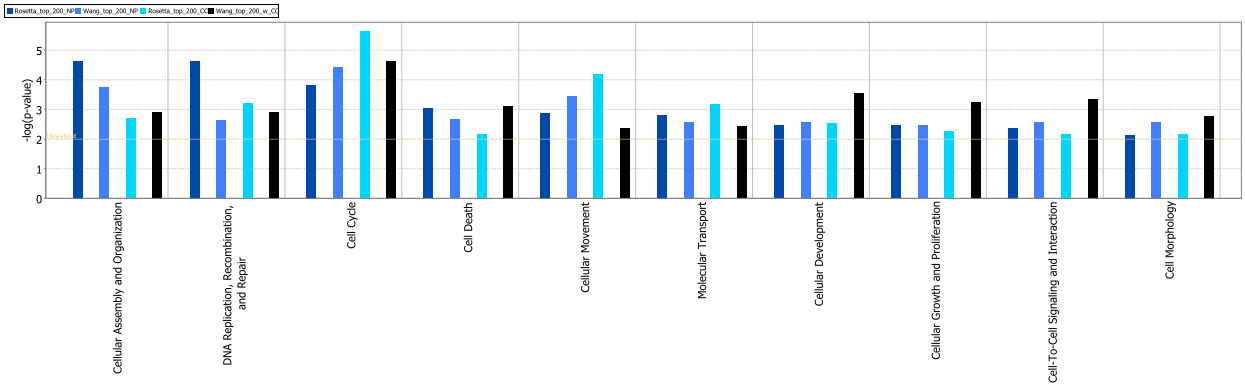
[8] H. Zhang, J. Ahn, X. Lin, and C. Park. Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, 22(1):88–95, 2006.



(A)Biological functions of top-200 selected genes by the Network Propagation



(B) Biological functions of top-200 selected genes by the Correlation Coefficients



(C) Biological functions of top-200 selected genes by the Network Propagation and the Correlation Coefficients

Figure 2: **Comparison of the identified biological functions enriched by the top-200 selected genes using Network Propagation and Correlation Coefficients in the Rosetta and Wang's datasets.** The functions are sorted by p-value calculated using the right-tailed Fisher Exact Test. Note that we only plot the functions which are involved with at least two genes on both datasets and have *p*-value less than 0.01. (A) The 15 over-represented functions enriched by the top-200 selected genes by Network Propagation. (B) The 12 over-represented functions enrighed by the top-200 selected genes by Correlation Coefficients. (C) The 10 common over-represented functions identified by Network Propagation and Correlation Coefficients.