Signed Network Propagation for Detecting Differential Gene Expressions and DNA Copy Number Variations

Wei Zhang
Department of Computer
Science and Engineering
University of Minnesota Twin Cities
Minneapolis, Minnesota
weizhang@cs.umn.edu

Nicholas Johnson
Department of Computer
Science and Engineering
University of Minnesota Twin Cities
Minneapolis, Minnesota
njohnson@cs.umn.edu

Baolin Wu
Division of Biostatistics
School of Public Health
University of Minnesota Twin Cities
Minneapolis, Minnesota
baolin@umn.edu

Rui Kuang*
Department of Computer
Science and Engineering
University of Minnesota Twin Cities
Minneapolis, Minnesota
kuang@cs.umn.edu

ABSTRACT

Network propagation algorithms have proved useful for the analysis of high-dimensional genomic data. One limitation is that the current formulation only allows network propagation on positively weighted graphs. In this paper, we explore two signed network propagation algorithms and general optimization frameworks for detecting differential gene expressions and DNA copy number variations (CNV). The proposed algorithms consider both positive and negative relations in graphs to model gene up/down-regulation or amplification/deletion CNV events. The first algorithm (Signed-NP) integrates gene co-expressions and differential expressions for consistent and robust gene selection from microarray datasets by propagation on gene correlation graphs. The second algorithm (Signed-NPBi) identifies gene or CNV markers by propagation on sample-feature bipartite graphs to capture bi-clusters between samples and genomic features. Large scale experiments on several microarray gene expression datasets and CNV datasets validate that Signed-NP and Signed-NPBi perform better classification of gene expression and CNV data than standard network propagation. The experiments also demonstrate that Signed-NP is capable of selecting genes that are more biologically interpretable and consistent across multiple datasets, and Signed-NPBi can detect hidden CNV patterns in bi-clusters by smoothing on correlations between adjacent probes.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation, feature evaluation and selection, pattern analysis*

General Terms

Algorithms

Keywords

Signed Network, Network Propagation, Gene Expression, DNA Copy Number Variation, Graph-based Learning

1. INTRODUCTION

Powered by high-throughput genomic technologies, it is now common practice to perform genome-scale experiments for measuring gene expressions, copy number variations (CNVs), single nucleotide polymorphisms, and other molecular information for cancer studies. Correlating these high-dimensional genomic features with cancer phenotypes as molecular signatures (biomarkers) can possibly improve prognosis and diagnosis over current clinical measures for risk assessment of patients [18, 4, 16]. The most widely used statistical methods to detect biomarkers are Pearson correlation coefficients [18] and hypothesis test methods such as student t-test. There are two major limitations of these popular methods. First, the genomic features are ranked by their individual correlation with the phenotype, and thus relations among features, for example co-expressed genes under certain conditions and adjacent probes involved in the same CNV events, are ignored. Second, usually all the samples are used to compute the correlation with phenotype, and thus biomarkers specific to only a subset of the samples are not detectable.

Network propagation is a graph-based learning algorithm [24] similar to PageRank used by Google. It has been shown that network propagation is capable of capturing the dependence among genomic features to detect correlated features as biomarkers [23, 21]. An efficient network propa-

[*] To whom correspondence should be addressed.

ACM-BCB'12, October 7-10, 2012, Orlando, FL, USA

gation algorithm on bipartite graphs was introduced to explore sample-feature bi-clusters for feature selection and cancer outcome classification [8]. In the network propagation regularization framework, a quadratic term with a normalized graph Laplacian matrix as hessian is combined with a square-loss on the predictions to explore the global graph structure for capturing correlation between all genomic features. The graph Laplacian matrix is only defined for positively weighted graphs which poses a significant limitation on the applicability of network propagation to the analysis of genomic data. For example, gene expression data could require a precise representation of up-regulated expression or down-regulated expression, and CNV data require a precise representation of amplification or deletion events. In these real computational biology problems, genomic data are represented by signed graphs to incorporate both positive and negative relations and thus the existing network propagation algorithms are not applicable.

To address the problem, we propose two signed network propagation algorithms and regularization frameworks for detecting differential gene expressions and DNA copy number variations. In the frameworks, we introduce signed graph Laplacians into network propagation. The first algorithm, Signed-NP, runs network propagation on a gene graph weighted by both positive and negative gene co-expressions for gene selection from gene expression datasets. Signed-NP integrates gene co-expressions and differential expressions to explore gene modules. The second algorithm, Signed-NPBi, runs network propagation on sample-feature bipartite graphs linked by both positive and negative features to identify gene or CNV markers. Signed-NPBi explores bi-clusters between patients and features to find biomarkers specific to subsets of patient samples.

2. METHOD

We first review network propagation in section 2.1. We then introduce signed network propagation (Signed-NP) in section 2.2 and its extension for propagation on bipartite graphs (Signed-NPBi) in section 2.3. In section 2.4 and section 2.5 we apply Signed-NP on gene correlation graphs for detecting differential gene expressions, and Signed-NPBi on sample-feature bipartite graphs for gene selection or CNV detection respectively.

2.1 Network Propagation

Let $G = (V, W^{(+)})$ denote an undirected graph with vertex set V and positive adjacency matrix $W^{(+)} \in \mathbb{R}^{+|V|\times|V|}$. In network propagation, the vertex set V is initialized by a vector y which is the $+1/-1$ label on training vertices and 0 on test vertices in binary classification. In the regularization framework proposed by [24], the objective is to learn a label assignment function $f : V \to \mathbb{R}$ to assign labels to the test vertices. The cost function is defined as follows,

$$\Omega(f) = \sum_{i,j} W_{ij}^{(+)}\Big(\frac{f_i}{\sqrt{D_{ii}^{(+)}}} - \frac{f_j}{\sqrt{D_{jj}^{(+)}}}\Big)^2 + \varrho\|f - y\|^2, \qquad (1)$$

where $D^{(+)}$ is a diagonal matrix with $D_{ii}^{(+)} = \sum_j W_{ij}^{(+)}$ and $\varrho \geq 0$ is a parameter to weight the two terms in the cost function. The first term in Eqn. (1) is the *smoothness constraint*, which encourages assigning similar labels to strongly connected vertices. The second term is the *fitting constraint*, which encourages consistency between predictions and training labels. The first term can be rewritten as

$$f'(I - (D^{(+)})^{-\frac{1}{2}} W^{(+)} (D^{(+)})^{-\frac{1}{2}})f,$$

where $I - (D^{(+)})^{-\frac{1}{2}} W^{(+)} (D^{(+)})^{-\frac{1}{2}}$ is the normalized graph Laplacian, which is positive semi-definite. Thus Eqn. (1) is a quadratic problem with a closed-form solution. However, when G is a signed graph the hessian matrix is not guaranteed to be positive semi-definite and thus the framework is not valid for network propagation anymore.

2.2 Signed Network Propagation

To allow both positive and negative edges for network propagation, we introduce signed graph Laplacian [10] into the regularization framework. Given a signed graph $G = (V, W)$ with vertices V and adjacency matrix $W \in \mathbb{R}^{|V|\times|V|}$. The cost function of the regularization framework is modified as follows,

$$\Omega(f) = \sum_{i,j} |W_{ij}|\Big(\frac{f_i}{\sqrt{D_{ii}}} - sgn(W_{ij})\frac{f_j}{\sqrt{D_{jj}}}\Big)^2 + \varrho\|f - y\|^2, \quad (2)$$

where $D_{ii} = \sum_j |W_{ij}|$. The first term in Eqn.(2) is the normalized signed graph Laplacian $I - S$, where $S = D^{-\frac{1}{2}} * W * D^{-\frac{1}{2}}$. It has been shown in [10] that the signed graph Laplacian is always positive semi-definite. The first cost term encourages assigning similar labels to vertices connected by positive edges and opposite labels to the vertices connected by negative edges. Empirically, the eigenvalues of S can be very small. For better performance in network propagation, we rescale S by dividing the largest eigenvalue such that S's eigenvalues are in the range $[-1, 1]$. Similar to the algorithm proposed by [24], the optimization framework in Eqn.(2) can be solved with an iterative label propagation algorithm,

$$f^t = (1 - \alpha)y + \alpha S f^{t-1}, \qquad (3)$$

where t denotes the propagation step and $\alpha = 1/(1 + \varrho)$. The parameter α balances the weights between initial label and network structure. The larger the α, the more we trust the network structure. This algorithm simply propagates labels among the neighbors in the graph. The algorithm will converge to the closed-form solution

$$f^* = (1 - \alpha)(I - \alpha S)^{-1} * y, \qquad (4)$$

where f^* assigns labels to the vertices.

2.3 Propagation on Signed Bipartite Graph

We next extend the framework in Eqn.(2) for signed bipartite graphs. Let $G = (V, U, E, W)$ denote a signed bipartite graph, where V and U represent two disjoint vertex sets, E is a set of weighted edges, and $W \in \mathbb{R}^{V\times U}$ is the wighted adjacency matrix. Each edge $(v, u) \in E$ connects two vertices v and u with weight W_{vu}. The initialization function y for the two vertex sets are denoted by $y(v)$ and $y(u)$. In this context, the cost function over $G = (V, U, E, W)$ is defined as

$$\begin{aligned}\Omega(f) \quad &= \quad 2\sum_{(v,u)\in E} |W_{vu}|\Big(\frac{f(v)}{\sqrt{D_{vv}}} - sgn(W_{vu})\frac{f(u)}{\sqrt{D'_{uu}}}\Big)^2 \\ &\quad + \varrho\|f(v) - y(v)\|^2 + \varrho\|f(u) - y(u)\|^2, \qquad (5)\end{aligned}$$

where $\varrho \geq 0$ is a parameter for balancing the cost terms, and D and D' are diagonal matrices with $D_{vv} = \sum_{u\in U} |w(v, u)|$

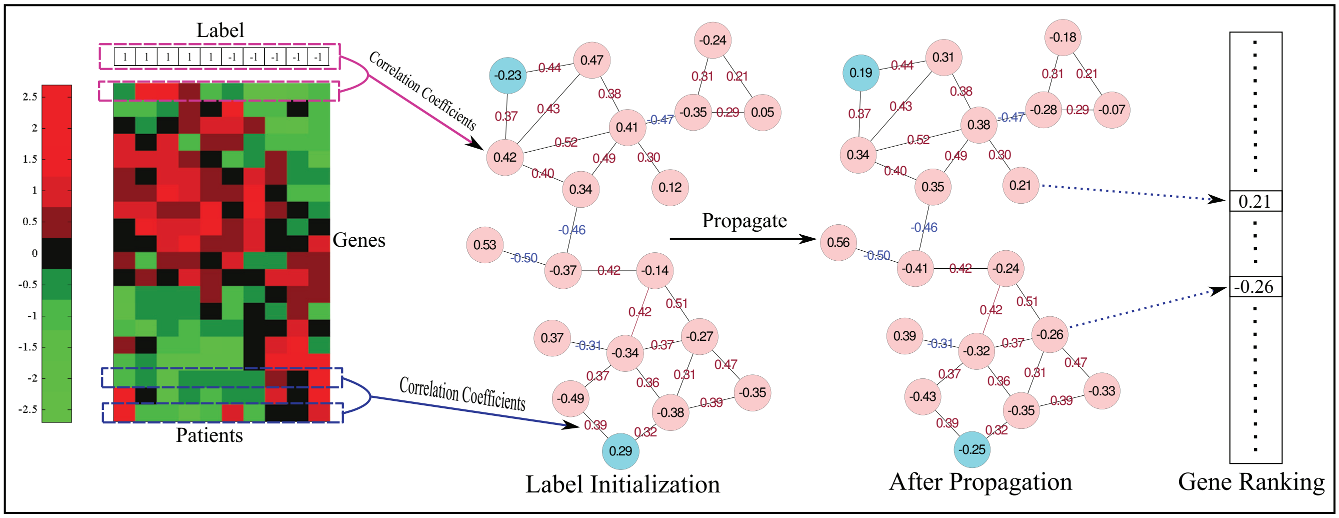


Figure 1: Running Signed-NP on a gene correlation graph. A gene correlation graph is constructed from gene expression data. The vertices are then initialized by the correlation between each individual gene expression and the labels. Network propagation on the signed graph re-rank all the genes for biomarker discovery.

and $D'_{uu} = \sum_{v \in V} |w(v,u)|$. The first cost term encourages similar labeling on positively connected vertex pairs and opposite labeling on negatively connected pairs. The second term and the third term constrain the new label assignment to be consistent with the initial labeling. To solve the optimization problem in Eqn.(5), we can also use a similar network propagation algorithm to compute the closed-form solution. The propagation algorithm iteratively performs propagation between the two vertex sets in both directions as follow,

$$f(v)^t = (1-\alpha)y(v) + \alpha S f(u)^{t-1}$$
$$f(u)^t = (1-\alpha)y(u) + \alpha S f(v)^{t-1}$$

where $\alpha = 1/(1+\varrho)$, $S = D^{-\frac{1}{2}} * W * D'^{-\frac{1}{2}}$, and t denotes the propagation step. S is also similarly rescaled by dividing the largest eigenvalue. Label information is propagated through neighbors in the bipartite graph. The algorithm will converge to the closed-form solution as in Eqn.(4).

2.4 Learning with Gene Correlation Graph

We first apply Signed-NP to a gene correlation graph for identifying differentially expressed genes. An illustrative example of network propagation on a gene correlation graph $G = (V, W)$ is shown in Figure 1. Each vertex in V represents a gene which is initialized by Pearson's correlation coefficients between gene expressions and the case/control labeling. An example of calculating the correlation between gene expression and labels is given in the pink rectangles in the figure. The initial labels provide the differential expression of each individual gene in the case/control study as the starting point of propagation. Each W_{ij} is the Pearson's correlation coefficients between gene expressions of gene i and gene j. An example of computing the correlation between gene expressions is given in the blue rectangles in the figure. Signed-NP propagates the initial labels across the network and the propagation process assigns similar labels to genes that are positively co-expressed and opposite labels to genes with opposite expression. The intuition is that we assume marker genes are active either in the case group or the control group but never both. Thus, the positive edges play the role to join positively co-expressed genes and the negative edges play the role to distinguish the genes with opposite expressions. After network propagation, the genes are re-ranked by the magnitude from positive to negative. Figure 1 shows how label propagation can capture the hidden clusters to recover false negatives and eliminate false positives. In the example, we assume two hidden clusters in the network, one of which contains a gene with initial value -0.23 and the another contains a gene with initial value 0.29 (the two cyan nodes). After running label propagation, final scores are assigned by balancing their coherence and discrimination so that genes in the same cluster are assigned similar scores.

2.5 Learning with Sample-Feature Bipartite Graph

We next apply Signed-NPBi to gene expression and CNV data for both genomic feature selection and sample classification. Gene expression data is modeled by a sample-feature bipartite graph $G = (V, U, E, W)$ as illustrated in Figure 2, where V represents the set of sample vertices and U represents the set of gene vertices. Each edge $(v, u) \in E$ connects sample vertex $v \in V$ and gene vertex $u \in U$ weighted by W_{vu} as is illustrated by the blue rectangles. The sample vertices in V are labeled with +1/-1/0 (case/control/unlabeled) as illustrated by the pink rectangle and the gene vertices are initialized with zeros. In this modeling, Signed-NPBi explores those bi-cluster composed of vertices with opposite labels and connected with negative edges as well as vertices with similar labels and connected with positive edges. As explained in the signed graph model in Eqn.(5), the first term in the cost function constrains new labeling to be consistent between the positively connected sample-gene pairs and opposite between the negatively connected sample-gene pairs. The second term is a fitting term which keeps the new label assignment for each sample consistent with the initial label. For the unlabeled vertices $v \in V$ with $y(v) = 0$, the second term is used to regularize these $f(v)$s such that the total cost is constrained. The third term is used in the same

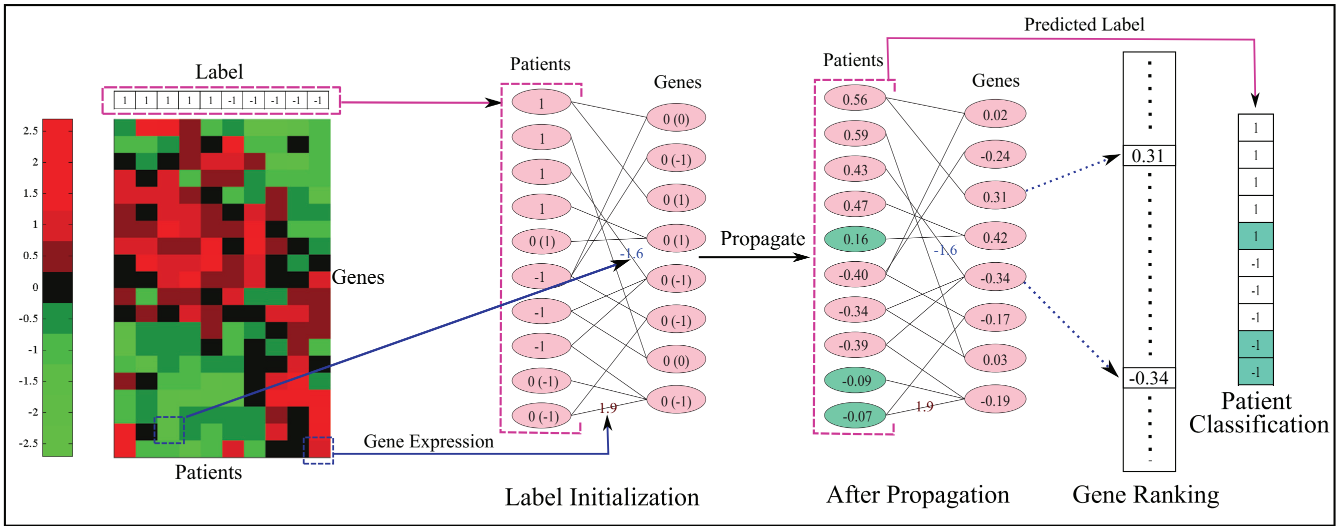


Figure 2: Running Signed-NPBi on sample-feature bipartite graph. Gene expression data is modeled as a sample-feature bipartite graph. The sample vertices are initialized by the case/control labels and the gene vertices are initialized with 0. Network propagation classifies the unlabeled samples and ranks the genes by their importance.

spirit to constrain the cost on the gene vertices. The final scores obtained after convergence are used to rank the genes as well as classify additional test (unlabeled) samples. In Figure 2 the optimal labels are given in parentheses. After running network propagation, the genes in bi-clusters will receive more significant values. Note that if connected by negative edges, the sample and the gene will receive opposite labels. Since the important genes are strongly connected to either the case group or the control group we can consider the genes with significant scores within the bi-clusters as biomarkers. The unlabeled samples are also classified to different groups based on the sign of the final score for each sample.

Similarly, we can also apply the signed bipartite graph model to copy number variation data. Each probe feature is represented by a vertex in U connected to the samples with an edge weighted by the log intensity ratio of the probe. After network propagation, those probes with high scores are selected as important CNV regions. In CNV analysis the adjacent probes tend to be strongly correlated, thus the bi-clusters in the bipartite graph represent a continuous CNV regions across a subset of samples.

3. EXPERIMENTS

In the experiments, we tested Signed-NP on 5 breast cancer gene expression datasets to detect differentially expressed genes and Signed-NPBi on two breast cancer gene expression datasets and one bladder cancer arrayCGH dataset to detect both differentially expressed genes and copy number variations.

3.1 Biomarker Identification from Gene Correlation Graph

GEO Index	GSE1456	GSE2034	GSE3494	GSE6532	GSE7390
Study	Pawitan	Wang	Miller	Loi	Desmedt
# of Meta	35	95	37	51	35
# of Meta-free	35	114	150	96	136

Table 1: Samples in five breast cancer datasets.

3.1.1 *Preparing breast cancer datasets*

We collected five independent microarray gene expression datasets generated for studying breast cancer metastasis. The five datasets were generated by the Affymetrix HG-U133A platform. The raw .CEL files were downloaded from the GEO website: Pawitan (GSE1456), Wang (GSE2034), Miller (GSE3494), Loi (GSE6532), and Desmedt (GSE7390) [13, 20, 12, 11, 3], and normalized by RMA [9]. After merging probes by gene symbols and removing probes with no gene symbol, a total of 13,261 unique genes derived from the 22,283 probes were included in our study. The patients are classified as cases and controls in the five datasets based on the time of developing distant metastasis. The patients who were free of metastasis for longer than eight years of survival and follow-up time were classified as metastasis-free and the patients who developed metastases within five years were classified as metastasis cases. The number of selected samples are reported in Table 1.

3.1.2 *Classification performance*

We applied Signed-NP, network propagation (NP), and Pearson's correlation coefficient (CC) to identify markers from each of the five breast cancer datasets. To test NP on a graph with positively weighted edges the network was constructed by setting the weights of the edges to the absolute value of the Pearson's correlation coefficients between the gene pairs and the vertices were initialized by the Pearson's correlation coefficients between gene expressions and the case/control labeling. To evaluate the predictive power of the marker genes we performed a cross-dataset validation. Specifically, we selected the markers genes for Signed-NP and CC from the top 50 up-regulated and 50 down-regulated genes and from the top 100 genes for NP from the training dataset, and evaluated the marker genes on the remaining datasets as test sets. The gene expressions of the marker genes were used as features for cross-validation on the test dataset. We evaluated the classification performance using a Support Vector Machine (SVM) [19] with an RBF kernel.

		Test Dataset				
Training Dataset	Method	GSE1456	GSE2034	GSE3494	GSE6532	GSE7390
GSE1456	Signed-NP		0.5985	0.6591	**0.6480**	**0.7247**
	NP		0.5827	0.6544	0.6424	0.6839
	Correlation Coefficients		**0.6019**	**0.6600**	0.6464	0.7070
GSE2034	Signed-NP	0.7830		**0.6183**	**0.7222**	**0.7471**
	NP	**0.7874**		0.6147	0.7186	0.7398
	Correlation Coefficients	0.7832		0.6174	0.7218	0.7361
GSE3494	Signed-NP	0.7940	**0.6410**		**0.6712**	**0.7492**
	NP	**0.7981**	0.6334		0.6699	0.7311
	Correlation Coefficients	0.7841	0.6165		0.6641	0.7209
GSE6532	Signed-NP	**0.7940**	**0.6332**	**0.6576**		**0.7380**
	NP	0.7867	0.6001	0.6409		0.6807
	Correlation Coefficients	0.7840	0.6298	0.6481		0.7006
GSE7390	Signed-NP	0.8103	**0.6357**	**0.6672**	0.6573	
	NP	0.8077	0.6177	0.6629	0.6510	
	Correlation Coefficients	**0.8150**	0.6232	0.6614	**0.6592**	
	Random	0.7475	0.6118	0.5883	0.6264	0.6229

Table 2: Evaluating marker genes across breast cancer datasets. Markers selected on the training dataset are used as features in the cross-validation on the test dataset. The best results are bold.

Classification performance was evaluated using a receiver operating characteristic (ROC) score [5]. We reported the mean of the ROC score after repeating 100 times five fold cross validation. We compared the predictive power of the marker genes selected by Signed-NP, NP, and CC. To add a random baseline, we also randomly selected 100 genes and tested with five-fold cross-validation 1000 times. The results are reported in Table 2 with $\alpha = 0.5$ for Signed-NP where the bold numbers represent the best ROC score by the three methods. Signed-NP outperformed both NP and CC in 14 out of 20 cases and NP alone in 18 cases. Signed-NP is clearly more capable of selecting more predictive marker genes in the experiments.

3.1.3 Consistency of marker genes across datasets

To measure how consistent the selected marker genes are across the five independent datasets, we report the percentage of common genes identified by a method in the rank lists from the datasets. This measurement assumes that the true marker genes are more likely to be selected in each dataset than the other genes. Thus, higher consistency across the datasets might indicate higher quality in gene marker selection. We plot the percentage of common genes among the first k (up to 1000) genes in the gene ranking lists from at least three of the datasets. We show the results of up-regulated genes in Figure 3. The network propagation method Signed-NP with parameter $\alpha = 0.5$ clearly identifies significantly more reproducible marker genes than CC. For example, Signed-NP identified 31 common genes among the first 100 genes in the gene ranking lists and CC only identified 24 common genes since CC only considers each feature independently. NP produced similar consistency in the marker genes compared with Signed-NP since both NP and Signed-NP capture the more conserved gene co-expressions.

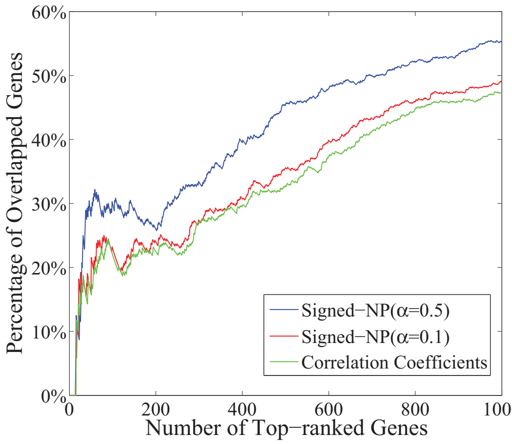
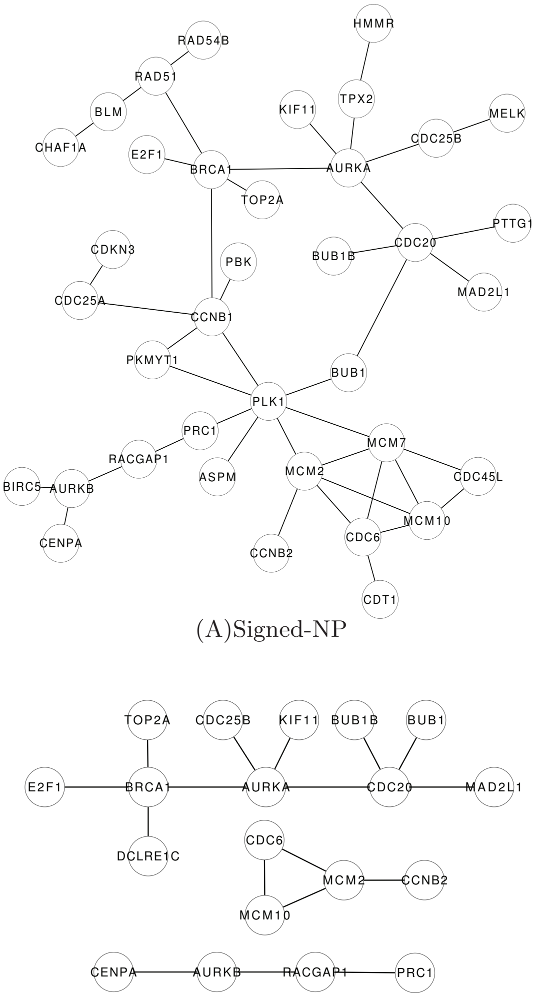


Figure 3: Marker gene consistency across five breast cancer gene list. The x-axis is the number of selected marker genes ranked by each method. The y-axis is the percentage of the overlapped genes between the selected markers across at least three of the breast cancer datasets.

3.1.4 PPI subnetworks and enriched GO terms.

The top 100 marker genes identified by Signed-NP and CC in the Desmedt (GSE7390) [3] dataset were mapped to the human protein-protein interaction (PPI) network obtained from HPRD [14] and also analyzed with the DAVID functional annotation tool [7]. We report the densely connected PPI subnetworks constructed from the top 100 up-regulated genes selected by Signed-NP in Figure 4(A). The subnetwork contains 37 genes and 43 connections between the genes. Compared with the PPI subnetwork generated from the marker genes selected by CC in Figure 4(B), which contain 19 genes and 17 connections, the subnetwork is larger and denser. In Figure 4(A), HMMR and RAD51 were reported as oncogenes of breast cancer in Online Mendelian Inheritance in Man (OMIM) [6], neither of which was detected by CC. Women with a variation in the HMMR gene had a higher risk of breast cancer even after accounting for mutations in the BRCA1 or BRCA2 genes. In particular, the risk of breast cancer in women under age 40 who carry the HMMR variation was 2.7 times higher than the risk in women without this variation [15]. RAD51 interacts with the evolutionarily conserved BRC motifs in the human breast cancer susceptibility gene BRCA2 [22]. In addition, the genes MAD2L1, RAD51, AURKA, BRCA1, BUB1, BUB1B, CDT1, and PTTG1 are listed on the breast cancer gene list in Genetic Association Database (GAD) [1]. Furthermore, the 37 marker genes in the subnetwork are also enriched by cell cycle process, nuclear division, DNA replication, DNA metabolic process, and ATP binding, all of which are well-known cancer relevant GO functions.



(A)Signed-NP

(B)Correlation Coefficients

Figure 4: Protein-Protein interaction subnetworks of signature genes identified by Signed-NP and Correlation Coefficients on the Desmedt dataset. (A) The PPI subnetworks identified by Signed-NP. (B) The PPI subnetworks identified by Correlation Coefficients.

The top 100 signature genes identified by Signed-NP enriched 83 GO functions (p-value<0.01) and the ones identified by CC only enriched 47 GO functions. The most significantly enriched GO functions are listed in Table 3. It is clear that Signed-NP identified signature genes that are more functionally coherent.

3.2 Genomic Feature Selection on Sample-feature Bipartite Graphs

3.2.1 Data Preparation

We prepared two microarray gene expression datasets [18, 17] to study breast cancer metastasis and one arrayCGH dataset to study bladder cancer [2]. The dataset (Rosetta) in [18] measures expression profiles of 24,481 genes generated by Agilent oligonucleotide Hu25K microarrays. This dataset contains 97 patient samples among which 51 patients were free of disease after their diagnosis for an interval of at least 5 years (good outcome) and 46 patients had developed distant metastasis within 5 years (poor outcome). The Vijver [17] dataset contains microarray gene expressions produced by the same technique for generating the Rosetta dataset on 295 samples (194 with good outcome and 101 with poor outcome). The two datasets were chosen for the experiment because Agilent array data by default report up/down-gene expression with positive and negative values for testing Signed-NPBi. The RMA normalized Affymetrix arrays used in the previous experiments usually contain absolute intensities. To avoid additional processing of the data, the five Affymetrix datasets were not used in this experiment. The arrayCGH dataset Blaveri [2] was generated with a HumanArray 2.0 array consisting of 2,464 probes at 1.5Mb resolution. After pruning, the dataset contained 98 samples and 2,142 probes. We classified the patient samples by the tumor stage.

GO terms	Signed-NP	Correlation Coefficients
cell cycle	38.382	21.419
cell cycle process	34.078	19.295
cell cycle phase	32.805	17.377
M phase	32.329	17.547
mitotic cell cycle	31.849	17.468
mitosis	27.777	15.157
nuclear division	27.777	15.157
M phase of mitotic cell cycle	27.534	14.994
organelle fission	27.236	14.794
cell division	19.805	12.449
organelle organization	19.072	13.072
spindle	18.723	12.557
microtubule cytoskeleton	14.743	X
chromosome	14.389	X
nuclear part	14.028	X
regulation of cell cycle	13.611	X
cellular component organization	13.389	X
DNA replication	12.282	X
intracellular non-membrane-bounded organelle	12.091	X
non-membrane-bounded organelle	12.091	X
intracellular organelle part	11.667	X
organelle part	11.530	X
condensed chromosome	10.774	X
nucleus	10.529	X
chromosomal part	10.397	X

Table 3: Enriched GO terms by the signature genes. The p-values in $-\log_{10}$ scale are shown for the enriched GO terms. A "X" denotes a p-value larger than 1×10^{-10}.

Signed-NPBi was compared against SVM with linear and RBF kernels and the bipartite network propagation algorithm (NPBi) [8]. To apply NPBi, each feature in the datasets was split into two features to represent the positive portion and the negative portion in the original features. The parameter α for both Signed-NPBi and NPBi was chosen from $\{0.95, 0.5, 0.1\}$ in the analysis.

3.2.2 Classification of gene expressions

Algorithm	Rosetta	Vijver
Signed-NPBi	**0.7374**	0.6682
NPBi	0.7290	0.6162
SVM(linear)	0.7072	0.6708
SVM(RBF)	0.7030	**0.6830**

Table 4: The mean AUC scores of classifying patients with good/poor prognosis in the Rosetta and Vijver gene expression datasets.

Signed-NPBi was tested on the two breast cancer gene expression datasets. We performed four-fold cross-validation on each of the two datasets with two folds for training, one fold for validation, and one fold for testing. We first initialized the patient labels in the validation fold to zero and combined the training folds to learn the model and tune the best regularization parameter α. The results on the test fold with the optimal α are reported to measure the prediction performance. We repeated the four-fold cross-validation 100 times on each dataset for Signed-NPBi, NPBi, and SVM (linear and RBF kernel) with the same setting. The mean AUC scores for classifying patients in the test fold are shown in Table 4. The results on the Rosetta dataset in Table 4 show that Signed-NPBi outperformed both NPBi and SVM with linear or RBF kernel. On the Vijver dataset, Signed-NPBi also performed better than NPBi. Although the SVMs get better classification performance on Vijver dataset, Signed-NPBi also performed reasonably well. The results support that network propagation on signed bipartite graphs improved classification over propagation on positively weighted graphs.

3.2.3 *Classification of CNV data*

Algorithm	AUC
Signed-NPBi	**0.8654**
NPBi	0.8306
SVM(linear)	0.8565
SVM(RBF)	0.8585

Table 5: The mean AUCs of classifying patients by tumor stage in the bladder cancer CNV dataset.

We then evaluated Signed-NPBi on the bladder cancer CNV dataset (Blaveri). The cross-validation setup in this experiment was the same as the setup in section 3.2.2. The mean AUCs are reported in Table 5. Signed-NPBi also outperformed NPBi and SVMs.

3.2.4 *Interpretation of CNV Detection with Signed-NPBi*

Finally, we evaluated how well Signed-NPBi smooths the CNV data in order to remove noise and identify bi-clusters. Signed-NPBi smooths the weighting of adjacent probes since the probes in proximity are more likely to be correlated. To demonstrate the smoothing effect we plot the weights of CNVs on chromosome 17 obtained by Signed-NPBi and Correlation Coefficients in the top half of Figure 5. In the region between probe index 60-135, CC and Signed-NPBi with $\alpha = 0.1$ detected low association with tumor stage while Signed-NPBi with $\alpha = 0.8$ and $\alpha = 0.95$ detected a negative association. By examining the probe log-intensity-ratios across the patients shown in the bottom half of Figure 5, we can confirm an amplification bi-cluster within the negative group in this region which was not captured by CC or Signed-NPBi with small α. This example shows the strength of Signed-NPBi to recover hidden bi-clusters in CNV data by taking into account the dependence between nearby probes.

4. CONCLUSION

In this paper, we present network propagation models on signed graphs for feature selection and classification in high-dimensional microarray gene expression and copy number variation data. Network propagation is a promising approach to explore modular structures such as clusters or bi-clusters hiding in high-dimensional data. The signed network propagation models are a useful and important generalization for modeling positive and negative relations in biological networks.

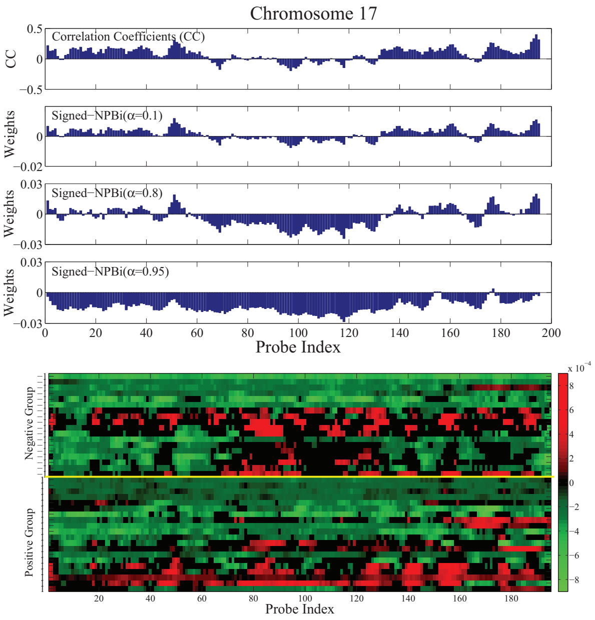


Figure 5: CNV weights learned by Signed-NPBi and Correlation Coefficients and the CNV data on Chromosome 17.

Since network propagation methods explore graph structures they are usually more computationally demanding compared with other simpler feature selection methods. Our future work will focus on developing approximations based on sparse structures to improve efficiency. In addition, we also plan to further investigate other regularizations of the signed graph Laplacian to improve the applicability and flexibility of the models.

5. ACKNOWLEDGMENTS

Funding: This work is supported by NSF grant III1117153, NIH GM083345, and NIH CA134848.

6. REFERENCES

[1] K. G. Becker, K. C. Barnes, T. J. Bright, et al. The genetic association database. *Nat Genet*, 36:431–432, 2004.

[2] E. Blaveri, J. L. Brewer, R. Roydasgupta, et al. Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clinical Cancer Research*, 11(19):7012–7022, 2005.

[3] C. Desmedt, F. Piette, S. Loi, et al. 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*, 13(11):3207, 2007.

[4] O. Gevaert, F. D. Smet, D. Timmerman, et al. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):184–190, 2006.

[5] M. Gribskov and N. L. Robinson. Use of receiver operating characteristic (ROC) analysis to evaluate

sequence matching. *Computers and Chemistry*, 20(1):25–33, 1996.

[6] A. Hamosh, A. F. Scott, J. S. Amberger, et al. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl 1):D514–D517, 2005.

[7] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[8] T. Hwang, H. Sicotte, Z. Tian, et al. Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics*, 24(18):2023–2029, 2008.

[9] R. Irizarry, B. Hobbs, F. Collin, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249, 2003.

[10] J. Kunegis, S. Schmidt, A. Lommatzsch, et al. Spectral analysis of signed graphs for clustering, prediction and visualization. *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM2010*, pages 559–570.

[11] S. Loi, B. Haibe-Kains, C. Desmedt, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of Clinical Oncology*, 25(10):1239–1246, 2007.

[12] L. Miller, J. Smeds, J. George, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*, 102(38):13550–13555, 2005.

[13] Y. Pawitan, J. Bjohle, L. Amler, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res*, 7(6):R953–R964, 2005.

[14] S. Peri, J. D. Navarro, R. Amanchy, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, 2003.

[15] M. Pujana, J. Han, L. Starita, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*, 39(11):1338–1349, 2007.

[16] T. R. Rebbeck, M. J. Khoury, and J. D. Potter. Genetic association studies of cancer: Where do we go from here? *Cancer Epidemiology Biomarkers and Prevention*, 16(5):864–865, 2007.

[17] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[18] L. van't Veer, H. Dai, M. Van de Vijver, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2001.

[19] V. Vapnik. Statistical learning theory. *NY Wiley*, 1998.

[20] Y. Wang, J. Klijn, Y. Zhang, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

[21] C. Winter, G. Kristiansen, S. Kersting, et al. Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol*, 8(5):e1002511, 05 2012.

[22] A. K. C. Wong, R. Pero, P. A. Ormonde, et al. Rad51 interacts with the evolutionarily conserved brc motifs in the human breast cancer susceptibility gene brca2. *Journal of Biological Chemistry*, 272(51):31941–31944, 1997.

[23] W. Zhang, B. Hwang, B. Wu, et al. Network propagation models for gene selection. In *Genomic Signal Processing and Statistics (GENSIPS), 2010 IEEE International Workshop on*, pages 1–4, nov. 2010.

[24] D. Zhou, O. Bousquet, T. Lal, et al. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16:321–328, 2004.

APPENDIX

A. SIGNED-NP

Using an iterative scheme, we find the closed optimal score which can minimize the cost function (2). For convergence during iteration, it is necessary to normalize W using diagonal matrix $D_{ii} = \sum_j |W_{ij}|$. A normalized matrix is calculated by $S = D^{-\frac{1}{2}} * W * D^{-\frac{1}{2}}$. The cost function (2) can be rewritten as,

$$\Omega(f) = f'(I - S)f + \varrho\|f - y\|^2.$$

With optimal function f^* for minimizing the cost function, the differentiated form is given by,

$$\frac{\partial\Omega}{\partial f} = 2(I - S) * f^* + 2\varrho(f^* - y) = 0.$$

This can be rewritten as the closed-form solution,

$$f^* = \frac{\varrho}{1+\varrho}(I - \frac{1}{1+\varrho}S) * y = (1-\alpha)(I - \alpha S)^{-1} * y,$$

where $\alpha = 1/(1+\varrho)$.

B. SIGNED-NPBI

Similar to the Signed-NP algorithm, we find the closed optimal score which can minimize the cost function (5) using the iteration scheme. We assign the new score by conserving the coherency among the connected genomic features and samples in the bi-cluster. To normalize the weight matrix two diagonal matrices are required: D_V with $D_{vv} = \sum_{u \in U} |w(v,u)|$ and D_U with $D'_{uu} = \sum_{v \in V} |w(v,u)|$. The normalized matrix is computed by

$$S = \left[\begin{array}{cc} 0 & D_V^{-\frac{1}{2}} * W * D_U^{-\frac{1}{2}} \\ D_U^{-\frac{1}{2}} * W^T * D_V^{-\frac{1}{2}} & 0 \end{array} \right],$$

The cost function (5) can be rewritten as follows,

$$\begin{aligned} \Omega(f) \quad &= \quad [f(v)^T \; f(u)^T] * (I - S) * \left[\begin{array}{c} f(v) \\ f(u) \end{array} \right] \\ &+ \varrho \left\| \left[\begin{array}{c} f(v) \\ f(u) \end{array} \right] - \left[\begin{array}{c} y(v) \\ y(u) \end{array} \right] \right\|^2, \end{aligned}$$

The solution f^* is identical to the one for Signed-NP.