# Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 07-021

## Identifying Clinical and Genetic Markers of Human Disease by Classifying Features on Graphs

Taehyun Hwang, Hugues Sicotte, Dennis Wigle, Jean-pierre Kocher,
Vipin Kumar, and Rui Kuang

September 26, 2007

# Identifying Clinical and Genetic Markers of Human Disease by Classifying Features on Graphs

TaeHyun Hwang[1], Hugues Sicotte[2], Dennis Wigle[3], Jean-Pierre Kocher[2], Vipin Kumar[1], and Rui Kuang[1*]

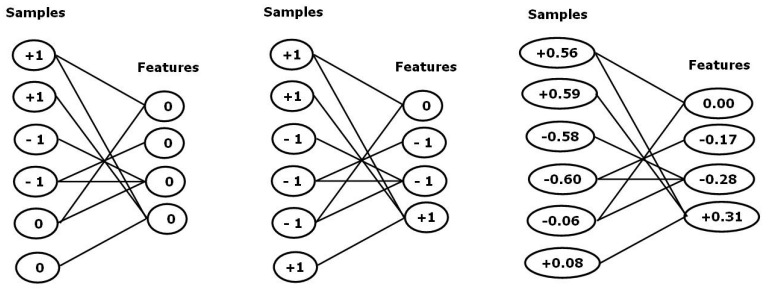
[1]Department of Computer Science and Engineering,
University of Minnesota, Twin Cities
[2]Bioinformatics Core, Mayo Clinic College of Medicine
[3]Division of General Thoracic Surgery, Mayo Clinic Cancer Center
[*]Corresponding author: kuang@cs.umn.edu

**Abstract.** Identification of clinical and genetic markers of disease can provide crucial information for both disease treatment and etiology. This complex task involves associating high-dimensional patterns such as large-scale gene expressions and single nucleotide polymorphisms (SNPs) with disease-related phenotypes using very few samples. We introduce a new graph-based semi-supervised *feature classification* algorithm to identify discriminative patterns by learning on bipartite graphs built from clinical variables, gene expressions and SNPs. Instead of performing feature selection or unsupervised bi-clustering, our algorithm directly classifies the feature nodes in a bipartite graph as positive, negative or neutral with network propagation, which captures the interactions between both samples and features (clinical and genetic variables) by exploring the global structure of the graph. Although globally optimized for classifying the features, our algorithm can also simultaneously classify the test samples for disease prognosis/diagnosis. We apply our algorithm to studying the Rosetta breast cancer dataset and chronic fatigue syndrome on a CAMDA contest dataset. Our algorithm identifies interesting clinical and genetic markers, some of which are consistent with previous studies in the literature, and achieves better overall classification performance than support vector machines and Bayesian networks.
(Supplemental website: http://compbio.cs.umn.edu/Feature_Class/.)

## 1 Introduction

Determining the causative factors of disease is critical for improving clinical treatment and understanding the biologic principles of disease. Recent developments in high-throughput technology allow large-scale measurement of genomic variations such as gene expressions and single nucleotide polymorphisms (SNPs) of a population. Associating these genomic variations with disease-related phenotypes provides good potential for elucidating the cause of disease [9]. However, computational identification of genetic markers of disease from high-throughput genomic data is an increasingly challenging problem. High-throughput data is

both expensive to generate and difficult to obtain. In a typical study, we only have tens of samples for analyzing several thousands of genes or even millions of SNPs, which suffers from the curse of "high-dimension and low sample size": the number of samples in every phenotype class is not large enough to represent the class distribution well. Using traditional biostatistics hypothesis testing or linear regression models leads to too many false positives and an intractable problem in variable selection. This curse of dimensionality also makes traditional machine learning strategies for feature selection [12] on high-throughput data particularly hard and unstable: to maximize the prediction performance of a classifier, existing algorithms rely on heuristics strategies for searching a sub-optimal feature set, which might not be unique given the high dimensionality of the feature set.

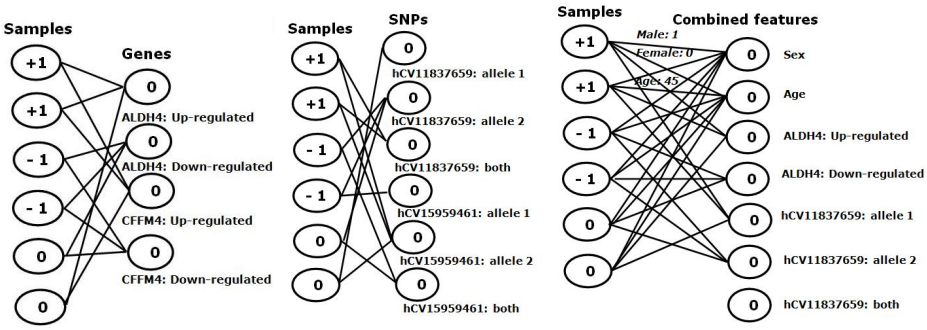



(A) Initial labels (B) Optimal labels (C) Predicted labels

**Fig. 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 classes, and 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.

In this paper, instead of selecting features based on their discriminative power to classify the samples, we propose to use the labeled samples to classify the features. We introduce a semi-supervised learning algorithm to associate integrated patterns of clinical variables, biomarker genes and SNPs with specific phenotypes in a disease context. We formulate the pattern-discovery task as a "hybrid" semi-supervised classification problem: we use training samples (positive and negative) to classify both the test samples and the features into positive, negative or neutral classes (Figure 1A and 1B). Here, the positively classified features and negatively classified features are candidate biomarkers. This new learning algorithm can capture the interactions between both samples and features (clinical and genetic variables) by exploring the global structure of the bipartite graph, based on a "cluster assumption": samples in the same class tend to be heavily

connected to a common set of features; features that can characterize a class tend to be heavily connected to the samples in the class.




(A) Gene Expressions (B) SNPs (C) Hybrid

**Fig. 2. Build bipartite graphs with clinical variables, gene expressions and SNPs.** (A) A bipartite graph with vertices of gene expressions; the edge weights are the absolute expression levels of the genes. (B) A bipartite graph with vertices of SNPs; all the edges are uniformly weighted by 1. (C) A bipartite graph with vertices of combined features. Different types of edges are rescaled with linear factors determined by cross-validation on the training set.

In the bipartite graph, *feature vertices* represent clinical variables, homozygous or heterozygous SNPs and up/down-regulated genes; *object vertices* represent labeled (+1/-1) and unlabeled (0) samples, connected to the feature vertices by weighted edges. Every clinical variable is denoted by one vertex, and it is connected to all the samples by the edges weighted by the original clinical values. Every gene is represented by two vertices, up-regulated or 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 2A). Each SNP will have three states, two homozygous states and one heterozygous state; every sample will be connected to one of the three vertices depending on the SNP type of this sample (Figure 2B). The bipartite graph can also contain heterogeneous vertices combined from clinical variables, gene expressions and SNPs (Figure 2C). 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.

Our algorithm is in a family of label propagation algorithms on graphs [19, 1], related to spectral graph theory, which can been regarded as semi-supervised graph-cutting techniques achieving the global optimal solution [1]. Recognized as having good generalization ability and excellent efficiency, these algorithms are receiving increasing attention from both machine learning and computational biology research communities [16, 14, 7]. The common property of all these graph-based learning algorithms is the "cluster assumption": there are often subtle underlying cluster structures in a large graph, which can be used implicitly to improve classification of unlabeled samples. We formulate the problem differently by classifying objects and features together. To our knowledge, the

semi-supervised graph-based learning algorithms have not been used for feature identification problems on high-throughput genomic data previously.

Using the phenotype information in a supervised manner to discover discriminative features, our semi-supervised algorithm should be distinguished from those unsupervised techniques such as clustering or bi-clustering. Bi-clustering [13] and other graph theoretical approaches [17] have previously been applied to identify genetic markers of disease. Not commonly used for sample classification, these algorithms typically focus on finding graph cliques or sub-patterns that are strongly consistent with the focused population with heuristic solutions, while in contrast, our algorithm uses phenotype labels to achieve a global optimum for classifying "positive", "negative" and "neutral" features along with test samples in one semi-supervised learning procedure. Many other supervised machine learning techniques have also been applied to identify clinical and genetic markers of disease. These approaches are typically variations of commonly used supervised learning algorithms, such as SVMs [18] or Bayesian Networks [4]. These algorithms, which are not directly designed for supervised pattern discovery, are not in the same category of our algorithm.

The nature of our graph-based learning algorithm captures correlations between all features simultaneously by exploring the global graph structure. Our method is essentially a non-linear method for selecting features. After relaxing the labels into real numbers, our method can always achieve the unique global optimum with an efficient network propagation algorithm. The time complexity of our algorithm scales linearly with the total number of features. Thus, our method is stable and fast even under the curse of dimensionality in bio-marker detection. Finally, our semi-supervised learning algorithm naturally uses unlabeled data in the process of classifying the features, which can possibly improve the quality of the selected features.

## 2 Method

In this section, we first define our formulation of marker discovery and disease diagnosis/prognosis as a semi-supervised learning problem on bipartite graphs. An efficient network propagation algorithm is then introduced to compute the closed-form solution of the objective function for the semi-supervised learning.

### 2.1 Semi-supervised Learning on Bipartite Graphs

Formally, we define an undirected bipartite graph $G = (V, U, E, w)$, where $V$ and $U$ are two disjoint vertex sets and $E \in V \times U$ is a set of weighted edges; each edge $(v,u) \in E$ connects two vertices $v$ and $u$ with a positive weight $w(v, u)$. Let $d(v) = \sum_{(v,u)\in E} w(v, u)$ and $d(u) = \sum_{(v,u)\in E} w(v, u)$ denote the sum of the weights of the edges on the same vertex. Let $y : V \rightarrow \{-1, 0, +1\}$ be the initialization function assigning initial labels to the labeled and unlabeled vertices in $V$ and $U$. Let $f$ denote a label-assignment function over vertex sets $V$ and $U$. If we let $V$ be the sample set and $U$ be the variables/feature set. A label assignment on a variable indicates its association with a sample class. Under this context, we

define an objective function over $G = (V, U, E, w)$ as follows,

$$\Omega(f) = \sum_{(v,u)\in E} w(v,u)(\frac{f(v)}{\sqrt{d(v)}} - \frac{f(u)}{\sqrt{d(u)}})^2$$
$$+\varrho \sum_{v\in V}(f(v) - y(v))^2 + \varrho \sum_{u\in U}(f(u) - y(u))^2, \quad (1)$$

where $\varrho > 0$ is a regularization parameter for balancing the cost terms on the right side of the equation. The first term enforces a consistency between the strongly connected vertex pairs $(u, v) \in V \times U$. This term penalizes those $f$ functions with a cost proportional to the $w(v, u)$ if $f$ assigns different labels to $v$ and $u$. The second term is a fitting term which keeps the new label assignment consistent with the initial labeling. This can be viewed as a supervised way of minimizing the training errors, which are measured by the difference between the initial labels $y(v)$ and the new label $f(v)$ for labeled vertices $v \in V$. 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 spirit to constrain the cost on the vertices in $U$.

If we restrict the labels to discrete values, i.e. $f : V \cup U \rightarrow \{-1, 0, +1\}$, minimizing $\Omega(f)$ is NP hard. But if we relax the label values as $f : V \cup U \rightarrow R$, $\Omega(f)$ is convex and differentiable. Let $D_U$ be a diagonal matrix with $D_{i_u i_u} = d(u)$ and $D_V$ be a diagonal matrix with $D_{i_v i_v} = d(v)$, where $v \in V$ and $u \in U$, and $i_v$ and $i_u$ are the index of vertices $u$ and $v$ in the matrix. We define the normalized connectivity matrix $S$ of $G$ as follows,

$$S = \begin{bmatrix} 0 & D_V^{-\frac{1}{2}} * W * D_U^{-\frac{1}{2}} \\ D_U^{-\frac{1}{2}} * W^T * D_V^{-\frac{1}{2}} & 0 \end{bmatrix}, \quad (2)$$

where $W$ denotes a $|V|$ by $|U|$ matrix with $W_{i_v, i_u} = w(v, u)$. The closed-form solution $f^*$ of $\Omega(f)$ can be computed by

$$f^* = (1 - \alpha)(I - \alpha S)^{-1} y, \quad (3)$$

where $\alpha = \frac{1}{1+\varrho}$ and $I$ is the identity matrix (See Appendix A for proof).

### 2.2 Network Propagation Algorithm

It is computationally intensive to compute the matrix inverse in Equation 3, when the graph $G$ is large and contains a lot of non-zero entries in $S$. We use a network propagation algorithm to compute the closed-from solution more efficiently. The propagation algorithm iteratively performs a diffusion operation between the two vertex sets in both directions. Theoretically, the diffusion process will finally converge to the closed-form solution $f^*$ defined in Equation 3. The network propagation algorithm is described as follows,

1. Normalize the bipartite graph by computing $B = D_V^{-\frac{1}{2}} * W * D_U^{-\frac{1}{2}}$.

2. Choose parameter $\alpha$ and perform a two direction propagation, until convergency ($t$ denotes the time step):
   - for each $v \in V$, $f(v)^t = (1-\alpha) * y(v) + \alpha * \sum_{u \in U} B_{i_v i_u} * f(u)^{t-1}$
   - for each $u \in U$, $f(u)^t = (1-\alpha) * y(u) + \alpha * \sum_{v \in V} B_{i_v i_u} * f(v)^{t-1}$
3. The sequence $f^t$ converges to its limit $f^*$ and $f^*$ gives the class labels on the unlabeled vertices in both $V$ and $U$.

This algorithm propagates the label information of a vertex to its neighbors in the other vertex set. This propagation process will leverage the activation values of the vertices in a densely connected neighborhood; in other words, if we assume that the vertices with the same label tend to be in the same clusters in the graph, the vertices in the same class will eventually converge to having similar values (same labels). This iterative propagation process was originally proposed to spread the activation values in a psychology network [11]. It is intuitively consistent with the definition of our objective function in Equation 1. We can show that this algorithm converges to a closed-form solution (Equation 3) of the objective function $\Omega(f)$ (See Appendix B for proof). In Figure 1C, we show the predictions of Network Propagation on a toy graph. Note in [7], a similar algorithm has been used for protein ranking, but the normalization of $S$ is different and there is no regularization framework. Finally, when different types of vertices are connected to the samples with edges weighted at different scales, we use cross-validation on the training set to learn positive constants to rescale the edges.

## 3 Experiments

We test the Network Propagation algorithm on two public datasets: a breast cancer dataset [3] and a chronic fatigue syndrome (CFS) dataset. The second dataset is used for CAMDA contest in 2006 and 2007 (http://camda.bioinfo.cipf.es/). For our convenience, we use "Rosetta" and "CAMDA" to refer to the two datasets respectively. We compare the classification performance of our algorithm against SVMs with RBF kernels [15] and Bayesian networks [4]. The classification performance of all methods are evaluated using the receiver operating characteristics (ROC) score: the normalized area under a curve plotting the number of true positives against the number of false positives by varying a threshold on the decision values [5]. We also report the identified clinical and genetic markers in the two studies, and compare them with previous findings in the literature. In all experiments with SVMs and Network Propagation, we do cross-validation on the training set to pick the best parameters and compute the ROC on the test set. We repeat the process 50 times and report the mean and the variance.

### 3.1 Rosetta Breast Cancer Dataset

The Rosetta data are collected from lymph-node-negative breast cancer patients to search for correlations between gene expression profiles and clinical outcome. The dataset is divided into a training set of 78 patients and a test set of 19 patients [3]. The microarray gene expression profile measures the expression levels of 24,481 genes. [3] has identified 231 genes strongly associated with disease

outcome using the correlation coefficient of the expression for every gene with disease outcome. We test both 24,481 and 231 gene expressions in our experiments. The details for quantization and normalization of scanned microarray images are described in [3]. The clinical data consists of 8 variables: age, estrogen receptor positive (ERp), progesterone receptor positive (PRp), tumor size, tumor grade, angioinvasion, lymphocytic infiltration and Brca1 mutation. No preprocessing is performed for the clinical variables.

We build bipartite graphs with all the training and test samples with clinical variables, gene expressions, and both of them together. In Table 1, we report the classification results on the three different bipartite graphs. The Network Propagation algorithm outperforms the SVMs and the Bayesian networks when clinical variables or hybrid features are used. In the case of using gene features alone, the network propagation algorithm performs slightly worse than the SVMs.

**Table 1. Classifying patients on the Rosetta Dataset.** This table shows the ROC scores of classifying patients with good/poor prognosis on the Rosetta Dataset with SVMs, Bayesian networks and Network Propagation.

| **Method** | 231 Genes | | Clinical Var. | | Hybrid | |
|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std |
| SVMs | **0.843** | 0.014 | 0.788 | 0.019 | 0.754 | 0.018 |
| Bayesian Networks | 0.750 | 0.073 | 0.751 | 0.086 | 0.793 | 0.068 |
| Network Propagation | 0.833 | 0.001 | **0.862** | 0.011 | **0.844** | 0.004 |

In Table 2, we list the top-ranked clinical features, gene features and hybrid signatures selected by Network Propagation by the absolute values of Z-scores calculated from the activation values. The top ranked genes and clinical variables are very consistent between the hybrid case and the case using 231 genes or clinical variables alone. We also find that the top-ranked features are insensitive to the $\alpha$ parameter. The top-2 clinical variables selected by Network Propagation are ERp (oestrogen receptor positive) and PRp (progesterone receptor positive). These two clinical variables are different from the variables identified by Bayesian networks [4] and I-RELIEF feature selection [12], in which tumor grade and angioinvasion are reported as the most important clinical variables. ERp and PRp are clinically important variables, as they signify hormonal involvement in tumor progression and determine chemosentivity [3]. Our top-ranked genes are surprisingly consistent with the marker genes identified by I-RELIEF feature selection: the top 3 genes, CEGP1, PRAME and AL080059, are exactly matched. More interestingly, the up-regulated and down-regulated states of CEGP1 and AL080059 are both in the top-10 list of the hybrid signature, which indicates that these two genes might play important roles in cancer metastasis.

In Figure 3, we plot the expression patterns of the gene features. Those gene features with the highest or lowest activation values (leftmost columns and rightmost columns) show similar expression patterns. The expression patterns of the two groups of patients are distinct from each other. To estimate a significance value for classifying the features, we run Network Propagation on graphs with the same structure but randomized labels, and repeat the process 2000 times.

**Table 2. Selected Clinical and Genetic Markers on the Rosetta Dataset.** The sign behind every feature name denotes if the feature is classified as associated with the positive class (good prognosis) or the negative class (poor prognosis).

| Rank | Genes | Clinical Variables | Hybrid |
|---|---|---|---|
| 1 | CEGP1 (Down)$^-$ | ERp$^+$ | ERp$^+$ |
| 2 | **PRAME (Up)**$^-$ | PRp$^+$ | PRp$^+$ |
| 3 | **AL080059 (Down)**$^+$ | Diameter$^-$ | Diameter$^-$ |
| 4 | **PRAME (Down)**$^+$ | Angioinvasion$^-$ | **PRAME (Up)**$^-$ |
| 5 | BIRC5 (Down)$^+$ | Lymphocytic Infiltrate$^-$ | CEGP1 (Down)$^-$ |
| 6 | SEC14L2 (Down)$^-$ | Grade$^-$ | **AL080059 (Down)**$^+$ |
| 7 | RRM2 (Down)$^+$ | Brc1mutation | Age$^+$ |
| 8 | CCNB2 (Down)$^+$ | Age$^+$ | SEC14L2 (Down)$^-$ |
| 9 | **AL080059 (Up)**$^-$ | | **PRAME (Down)**$^+$ |
| 10 | Contig55725_RC (Down)$^+$ | | **AL080059 (Up)**$^-$ |

We estimate significance cutoffs for both the positive class and the negative class with a confidence level of 1% using all the 2000 activation values on all the feature nodes. We also compare the distributions on each feature nodes computed from activation values of the 2000 random runs with correlation coefficients. We observe clear block structures: distributions are very similar within positive gene features or negative gene features but not between the two groups. This suggests that our algorithm clearly identified gene features in the positive class and the negative class.

To compare the gene ranking induced by Network Propagation and correlation coefficients [3], we plot the rank of each gene in both rank lists against each other in Figure 4 in Appendix C. Clearly, majority of the genes are ranked differently by the two approaches, which suggests that many genes do not act independently and they often play joint roles with some other genes. It will be interesting to investigate further why those genes with low correlation coefficients are ranked much higher by Network Propagation. We are current under investigation of the functional roles of these genes.

### 3.2 CAMDA Dataset

In the CAMDA dataset, we are interested in classifying Ever CFS patients against Non-fatigue (NF) patients. The 78 patients (39 CFS versus 39 NF) with full data are used in our experiments. We compile a list of features with 38 clinical variables without missing values, 42 SNPs associated with 10 genes and 2,685 genes with high signal/noise ratio across all the samples.

The results of classifying CFS versus NF patients are shown in Table 3. When all SNPs are used for classification, the ROC score is around random. We decide to only use SNPs associated with NR3C1 as features and achieves improved results. The Network Propagation algorithm outperforms SVMs except for the hybrid case. The combined features do not produce improved results. The relatively high scores for clinical variables suggest that the clinicians assigned a patient status based on the clinical variable in a consistent way. Although our results show promising evidence of improvement over previous works in CAMDA

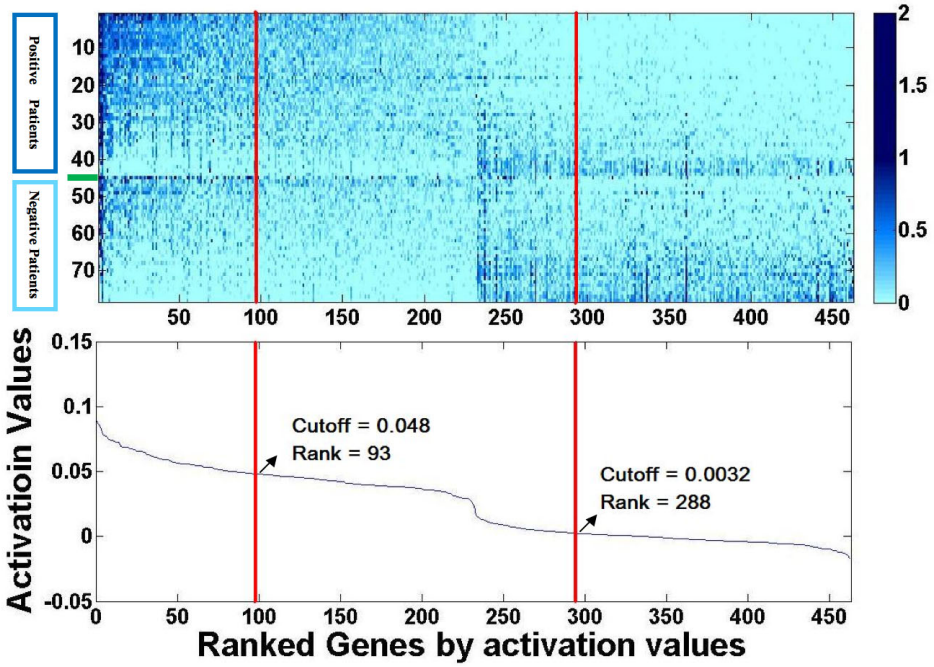



**Fig. 3. Expression patterns of the gene features ranked by activation values.** Two figures are aligned together to show the expression pattern of positive gene features, neutral gene features and negative gene features. In both figures, the columns correspond to gene features ranked by the activation values; in the upper figure, the rows correspond to patients ranked by the activation values after running Network Propagation, and in the lower figure, the rows correspond to the activation values of the gene features. Note each gene has two rows (up-regulated and down-regulated) in the figure; thus, all the expression levels are positive. The two red lines denote the estimated cutoffs of the significantly high and significantly low values.

contest in 2006, CFS is notoriously hard to diagnose and the patient group may have a heterogeneous set of molecular causes to their CFS symptoms.

On clinical features, the Network Propagation algorithm ranks several MFI variables as highly positive features and several SF-36 variables highly negative features, which is consistent with the statistical analysis in [10]: the symptoms of CFS patients that have been shown to have low scores in the SF-36 indicate more severe condition, whereas this is indicated by high scores in the MFI. Among the top-ranked genes, DKFZP434O047, HIPK2 and CEBPA have both the up-regulated and down-regulated states within first 100 gene features. While little is known about DKFZP434O047 and KBTBD9, the other genes all seem to have functions compatible with current CFS hypotheses with support in the literature: GnRH2 (gonadotropin releasing hormone 2) is found under expressed in patients with part of their pituar gland removed [6, 2]; RERE links to neurodegenerative disorders [6] and the other genes all seem to play a role in tumor suppression, cell proliferation, and cell division in tissue differentiation which are compatible

with T-cells activation [6]. The SNP markers of the glucocorticoid receptor gene NR3C1 are highly ranked by Network Propagation, which supports that the NR3C1 is strongly associated with CFS by playing a major role in the immune system and in the hypothalamicpituitary-adrenal (HPA) axis activity related to the immune function. Our top-ranked SNP markers such as rs2918419, rs860458, and rs6188 are consistent with the SNP markers identified by [8].

**Table 3. Classifying patients on the CAMDA Dataset.** This table shows the ROC scores of classifying CFS and NF patients on the CAMDA Dataset with SVMs and Network Propagation.

| **Method** | Genes | | Clinical Var. | | SNPs (NR3C1) | | Gene+SNPs | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SVMs | 0.647 | 0.154 | 0.842 | 0.019 | 0.590 | 0.143 | **0.669** | 0.148 |
| Network Propagation | **0.667** | 0.153 | **0.996** | 0.011 | **0.670** | 0.130 | 0.640 | 0.124 |

## 4 Discussion

We present a new formulation of supervised bio-marker selection: instead of selecting a set of features to maximize the predictive power or finding sub-patterns that are consistent with patient labels, we classify the features into "positive", "negative" and "neutral" groups. We also design an efficient semi-supervised graph-based learning algorithm to compute the global optimal solution of this feature classification problem.

One limitation of our current algorithm is that no prior knowledge about the dependence between the feature nodes such as linkage disequilibrium between SNPs, is used in the process of classifying features. We plan to design new algorithms that can utilize the dependence between the features as prior by running network propagation on graphs of more sophisticated topologies. Another challenge of applying the network propagation algorithm to this study is the integration of heterogeneous feature vertices in one graph. The strategy of learning linear weighting of edges with cross-validation poses a computational issue. Despite this fact, our approach is promising enough to warrant the design of new frameworks for a unified learning of both the feature weights and the scaling factors. Finally, although the network propagation algorithm demonstrates high prediction performance, no improvement has been achieved from data integration. We postulate that naive linear concatenation of different types of features is not a principled approach for data integration and possibly, putting different feature nodes in one graph is not the right strategy for the network propagation algorithm. Thus, we are also looking into algorithms that can integrate different graphs in a non-linear manner. As the volume and complexity of integrated genomic data continue to increase, we look forward to applying our graph-based learning algorithms to discovery-oriented problems with clinical implication.

## Acknowledgments

This work is supported by the Biomedical Informatics and Computational Biology Seed Grant for UM-Mayo-IBM Collaboration. Minnesota Super-computing Institution provided the computational facility for the work in this paper.

# Appendix

## A The closed-form solution

Similar to the derivation in [19], we can rewrite the Equation 1 as follows:

$$\begin{aligned}\Omega(f) &= \sum_{(u,v)\in E} w(v,u)(\frac{f(v)}{\sqrt{d(v)}} - \frac{f(u)}{\sqrt{d(u)}})^2 + \varrho \sum_{v\in V}(f(v)-y(v))^2 + \varrho \sum_{u\in U} f(u)^2 \\ &= [f(V)^T f(U)^T] * (I-S) * \begin{bmatrix} f(V) \\ f(U) \end{bmatrix} + \varrho || \begin{bmatrix} f(V) \\ f(U) \end{bmatrix} - \begin{bmatrix} y(V) \\ y(U) \end{bmatrix} ||^2\end{aligned}$$

Now, we differentiate $\Omega(f)$ with respect to $f$,

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

which can be arranged into

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

## B Convergency of network propagation

We can rewrite the network diffusion algorithm in matrix form as

$$\begin{aligned}f(V)^t &= (1-\alpha)y(V) + \alpha B * f(U)^{t-1} \\ f(U)^t &= (1-\alpha)y(U) + \alpha B^T * f(V)^{t-1},\end{aligned}$$

which can be rearranged as $f^t = (1-\alpha)y + \alpha S * f^{t-1}$. Following the proof in [19], we can show that $f^* = (1-\alpha)(I-\alpha S)^{-1}y$, which is exactly the closed form solution of Equation 1.

## C Comparison of the gene ranking by Network Propagation and Correlation Coefficients (Figure 4)

## D Ranking of the features on CAMDA dataset (Table 4)

## References


1. Y. Bengio, O. Delalleau, and N. L. Roux. Label propagation and quadratic criterion. In Eds. O. Chapelle, B. Schlkopf, and A. Zien, editors, *Semi-Supervised Learning*. MIT Press, 2006.
2. O. M. Dekkers, A. A. Klaauw, A. M. Pereira, N. R. Biermasz, P. J. Honkoop, F. Roelfsema, J. W. Smit, and J. A. Romijn. Quality of life is decreased after treatment for nonfunctioning pituitary macroadenoma. *Obstetrical and Gynecological Survey*, 62(1):33–34, 2007.
3. Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.
4. 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.

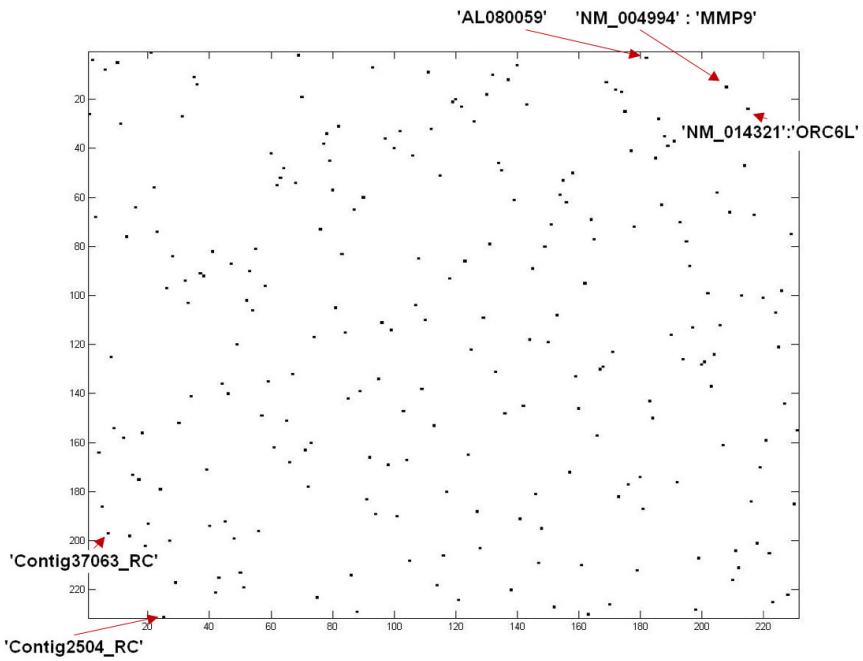



**Fig. 4. Comparison of the gene ranking by Network Propagation and Correlation Coefficients.** Rows correspond to the gene ranks by Network Propagation and columns correspond to the gene ranks by correlation coefficients. Each dot at $x$th row and $y$th column in the plot corresponds to a gene ranked at $x$th by Network Propagation and $y$th by correlation coefficients. Those genes with similar ranks will be close to the diagonal.


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. CDC CFS Group. http://www.cdc.gov/cfs/cfscauses.htm, 2007.
7. R. Kuang, J. Weston, W. S. Noble, and C. Leslie. Motif-based protein ranking by network propagation. *Bioinformatics*, 21(19):3711–3718, 2005. In press.
8. M. S. Rajeevan, A. K. Smith, I. Dimulescu, E. R. Unger, S. D. Vernon, C. Heim, and W. C. Reeves. Glucocorticoid receptor polymorphisms and haplotypes associated with chronic fatigue syndrome. *Genes, Brain and Behavior*, 6(2):167–76, 2007.
9. T. R. Rebbeck, M. J. Khoury, and J. D. Potter. Cenetic association studies of cancer: Where do we go from here? *Cancer Epidemiol Biomarkers Prev*, 16(5):864–5, 2007.
10. W. C. Reeves, D. Wagner, R. Nisenbaum, J. F. Jones, B. Gurbaxani, L. Solomon, D. A. Papanicolaou, E. R. Unger, S. D. Vernon, and C. Heim. Chronic fatigue sydrome - a clinically empirical approach to its defninition and study. *BNC Medicine*, 3(19), 2005.
11. J. Shrager, T. Hogg, and B. A. Huberman. Observation of phase transitions in spreading activation networks. *Science*, 236:1092–1094, 1987.

**Table 4. Selected Clinical, Genetic and SNPs Markers on the CAMDA dataset.** The sign after every feature name denotes if the feature is classified as related to positive class (CSF) or negative class (Non-fatigue). In the table, "A1" and "A2" denote the two homozygous SNP states.

| Rank | Genes | Clinical Variables | SNPs (Associated Gene) |
|---|---|---|---|
| 1 | DKFZP434O047 (Down)$^+$ | Physic$^-$ | rs2918419-Both (NR3C1)$^-$ |
| 2 | GNRH2 (Down)$^+$ | Vitality$^-$ | rs860458-Both (NR3C1)$^-$ |
| 3 | TP73 (Down)$^+$ | Social Funct$^-$ | rs933271-A2 (COMT)$^-$ |
| 4 | KBTBD9 (Down)$^+$ | Activ Reduc$^+$ | hCV7911132-A2 (SLC6A4)$^-$ |
| 5 | HCP5 (Down)$^+$ | Gen Fat$^+$ | rs258750-A1 (NR3C1)$^+$ |
| 6 | PDCD8 (Down)$^-$ | Bodily Pain$^-$ | rs1396862-A1(CRHR1)$^-$ |
| 7 | DOK1 (Down)$^+$ | Mental Fat$^+$ | rs6196-A2 (NR3C1)$^+$ |
| 8 | RERE (Down)$^-$ | Phys Fat$^+$ | rs6196-Both (NR3C1)$^-$ |
| 9 | RHOD (Down)$^+$ | Role Emotional$^-$ | rs2070762-A1 (TH)$^+$ |
| 10 | HIPK2 (Down)$^+$ | Motiv Reduc$^+$ | rs2066713-Both (SLC6A4)$^+$ |


12. Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23(1):30–37, 2007.
13. Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:S136–S144, 2002.
14. K. Tsuda, H. Shin, and B. Schölkopf. Fast protein classification with multiple networks. *bioinformatics*, 21:ii59–ii65, 2005.
15. V. N. Vapnik. *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, 1998.
16. J. Weston, A. Elisseeff, D. Zhou, C. Leslie, and W. S. Noble. Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences*, 101(17):6559–63, 2004.
17. N. Yosef, Z. Yakhini, A. Tsalenko, V. Kristensen, and A. Børresen-Dale. A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data. *Bioinformatics*, 23:e91–e98, 2006.
18. 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.
19. D. Zhou, J. Weston, A.Gretton, O.Bousquet, and B.Schoelkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 2004.