UNIVERSITY OF CALIFORNIA, SAN DIEGO

**Challenges in network-based classification of gene expression profiles**

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Sanath Kumar Ramesh

Committee in charge:

Professor Trey Ideker, Chair
Professor Vineet Bafna
Professor Charles Elkan

2012

The thesis of Sanath Kumar Ramesh is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2012

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Many people have helped me come this long in my graduate education. It is my pleasure to convey my gratitude to each and every one of them who have provided their assistance through this journey.

First and foremost, I offer my sincere gratitudes to my advisor, Prof. Trey Ideker for his supervision and guidance. Even though I lacked prior bioinformatics background, he gave me an opportunity to be a part of his research group and provided constant encouragement to keep trying. His assistance, willingness to discuss ideas, valuable comments and his confidence in me formed the backbone of this research.

My special thanks to Matan Hofree, fellow lab member for guiding me through the course of this thesis. His inputs and valuable insights on the fundamentals relevant to the subject were crucial to the successful completion of this thesis. I would also like to thank staff members and fellow lab students for providing a friendly and cheerful atmosphere to work.

Finally, I would like to thank my thesis committee members, Prof.Vineet Bafna and Prof. Charles Elkan for spending their precious time to provide critical comments which were vital for the successful realization of this thesis.

ABSTRACT OF THE THESIS

**Challenges in network-based classification of gene expression profiles**

by

Sanath Kumar Ramesh

Master of Science in Computer Science

University of California, San Diego, 2012

Professor Trey Ideker, Chair

Classification of gene expression profiles to distinguish one disease state from another is essential for the realization of personalized medicine. Recent approaches towards this problem use prior knowledge about interaction among biomolecules to improve classification accuracy and the biological relevance of the predictive features. However, many such network-based methods do not significantly outperform their unconstrained counterparts in terms of sensitivity and specificity due to unexplained reasons. This behavior, observed across diverse datasets and methods, is a cause of concern as it implies that something is wrong with the data, the algorithms or both. This work focuses on understanding the reasons behind this problem through extensive simulation of gene expression profile to help

the development of better classifiers in the future. We infer that when using networks whose interactions do not agree well with the patterns of gene expression, improvement in classification performance will not be significant. Because this improvement is dependent on the classifier also, future network-based methods need to understand their properties with respect to network noise and know the quality of actual network mapping to make meaningful inferences from the performance results.

# Chapter 1

# Introduction

Classification of tumor types for effective prognosis and treatment has been primarily based on histopathological appearances of the tissue. However, considering the diversity in treatment outcomes even within the same tumor class, many researchers have turned to genome-wide expression profiles to identify stronger prognostic and predictive markers for the disease. Some major challenges with biomarkers have been lack of reproducibility across cohorts [8] and, for some diseases, lack of classification accuracy [3]. This issue stems, in part, from the fact that not all the cases are the same disease, at least on the molecular level.

To address these shortcomings, many groups are beginning to implement classification approaches that draw from prior knowledge about cellular architecture and function to tie together individual genes and proteins into networks. Chuang et al. [3] highlighted the predictive power of protein subnetworks as opposed to individual gene markers for classification of breast cancer metastasis. In their approach nodes of the network are scored using the genes expression value and a greedy search is performed to identify high scoring subnetworks. Scores of such subnetworks are used as feature values for classification using a logistic regression framework. At about the same time, Rappaport et al. [18] also came up with a technique to include network information into support vector machines based on spectral graph theory. Following up these efforts, many methods [16] [24] [13] [15] [10] [6] [23] have been proposed to improve both the classification accuracy and the relevance of biomarkers.

Network biomarkers are beginning to improve diagnosis and stratification of disease [6] due to several key benefits. The two important ones are robustness across cohorts (Figure 1A) and enrichment for disease genes (Figure 1B). When similar sets of biomarkers are identified across cohorts, they tend to be reliable and more relevant for prognosis. Enrichment for disease genes argues that at least some network biomarkers identified are related to the causes of disease rather than some possibly distal downstream effect.

Puzzlingly, however, few network-constrained classification methods actually perform better than their unconstrained counterparts in terms of sensitivity and specificity (Figure 1C). On the one hand, this is not surprising as any feature combination available to the network-constrained method is also available when this constraint is removed and, in fact, the unconstrained method can access many additional feature combinations which may lead to better classification performance. On the other hand, it is conceivable that at least some network-constrained method may outperform an unconstrained classifier since it can better generalize across cohorts.

Therefore, it is yet unclear why network-based methods do not outperform network-free methods in classification of patient molecular profiles. Potential reasons could be a. Deficiency in network knowledge i.e. interactions in the network either wrongly or partially capture the expression pattern of genes. b. The network-based methodologies proposed to-date are unable to effectively use information encoded by the biological network. c. Patient molecular profiles are simply not as complex as widely assumed i.e., some disease states are already well-captured by a linear combination of a small number of genes, meaning that prior network knowledge is simply unnecessary for good classification. This leads us to a more immediately pressing question: are there certain data sets / classification problems for which we expect network-based biomarkers to outperform a standard biomarker set based on individual mRNA or protein levels?

To answer some of these questions, we analyze the performance of network-constrained classification via an extensive set of simulations. We explore the influence of key factors on network-based classification: a. Coverage and error of

network mapping; and b. Size of the disease pathway influencing a disease. We show that under certain conditions, increasing network coverage and accuracy can in fact strongly improve the performance of molecular classification. Finally, we highlight that contrary to popular belief, classification performance alone is only a partial metric for assessing the utility of network information. A complete understanding of the properties of classifiers with respect to mapping error and knowing the quality of the actual network mapping are essential for a realistic assessment.

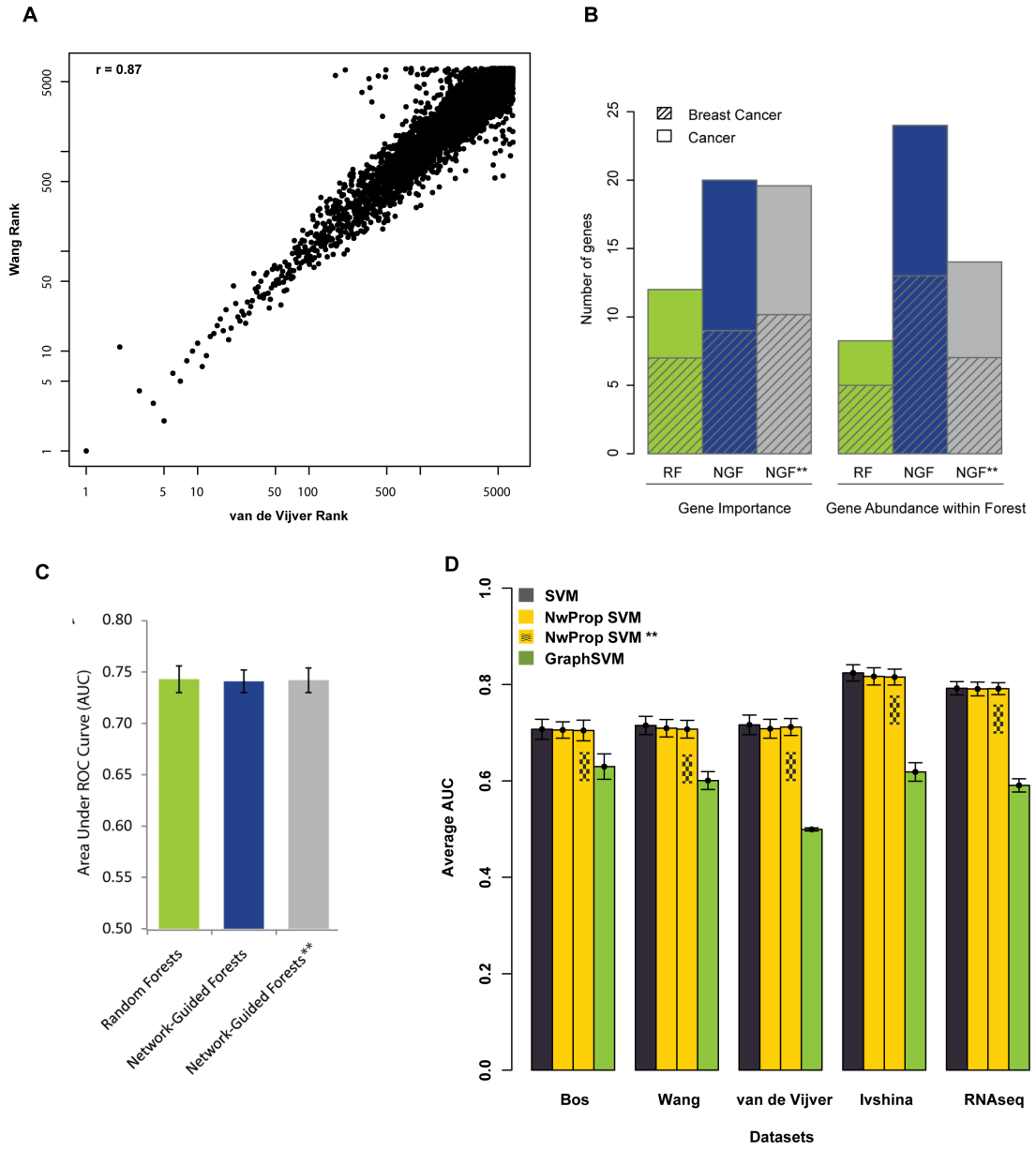


**Figure 1**: **Performance of network-based classifiers on real datasets.** (A) Correlation of the features ranked by NwpropSVM on two different datasets show that the results tend to be highly reproducible across cohorts. (B) and (C) are reproduction of Figures 4B and 4C from Dutkowski et al. (B) Among the top 25 features selected by Network Guided Forest (NGF), Random Forest (RF) and NGF on a permuted network (NGF **), NGF selects a significant amount of disease related genes. (C) Classification performance of NGF in terms of AUC does not improve when compared to RF and NGF ** (D) Other network-based classifiers exhibit similar behavior as in Panel C on five different expression datasets. ** denotes that the algorithm was run on a permuted network.

# Chapter 2

# Materials and Methods

## 2.1 Overview of the simulation framework

The purpose of simulation is to create an expression dataset where there is full control over the characteristics of the data while being able to implant a signal that is representative of some the biological properties found in gene expression profiles. Therefore, a protein-protein interaction network, downloaded from the STRING [20] database is used as a prior knowledge to generate patient molecular profiles that follow the interaction structure (Figure 2). Each profile is assigned a phenotype which is typically binary such as "disease" and "normal". Patients that have a disease show increased or reduced expression for certain disease genes whose expression value is drawn from a normal distribution with mean $\mu$ and standard deviation $\sigma$. All other expression values are drawn from a standard normal distribution. Since the parameters $\mu$ and $\sigma$ control the signal to noise ratio in the data, their values are empirically determined to minimize signal content as described in the results chapter.

To encode prior knowledge of interactions into the expression data, disease genes are selected from the pathway structures present in the network. This idea is motivated by the observation from many expression studies that genes causal to a disease commonly fall within pathways that are responsible for some cellular function such as cell cycle, DNA repair etc. Also, these genes within a pathway are known to interact heavily among themselves than with genes from other pathways

giving raise to an interaction topology called communities. To make the simulation generic, instead of choosing disease genes from biologically validated pathways, we choose them from regions of an interaction network that exhibit the community structure. Even though a pathway could be held responsible for a disease, not all genes in the pathway get expressed in all disease patients. This stochasticity is the result of widely varying genetic, environmental and demographic makeup of each patient and is also one source of complexity within expression datasets. This is precisely the reason why we expect network-based classifiers to outperform their unconstrained counterparts, as they can grasp the common expression pattern within pathways instead of getting confounded by the stochasticity in individual gene expressions. Therefore, in the simulation also, every disease gene gets differentially expressed in a random subset of disease patients. We add noise to the simulation by differentially expressing a few genes among randomly chosen normal patients.

To generate the expression data, pathways were identified in the STRING network by detecting network communities using the Qcut algorithm [19]. For simulations, a 2000 node network was constructed out of the original network by selecting pathways until enough number of nodes is present. Results presented in this work (Figure 3A and Figure 4) were generated with one pathway of around 50 genes showing differential expressed in disease patients. The generated expression data has 600 patients split into 300 disease and 300 normal patients. Equal number of disease and normal patients was fixed to avoid classifier bias towards the over-represented class. To encode the stochasticity in expression, each disease gene was differentially expressed in 20% of the disease patients picked at random. Increasing the size of this subset would reduce the stochasticity, and hence the usefulness of a network-based classifiers. Expression data that is synthesized using this approach is classified using three different classifiers  a regular support vector machine (SVM) [4] that does not use network information; GraphSVM [18] and NwpropSVM (See Section 2.3) that are extensions to SVM made to include a network into classification. Both network-based methods are supplied with the original network which was used to generate the expression data and their classi-

fication performance is compared against SVM as a baseline.

Admittedly, this model is a clear simplification of the true biology by making use of only few essential characteristics that have been established about biological

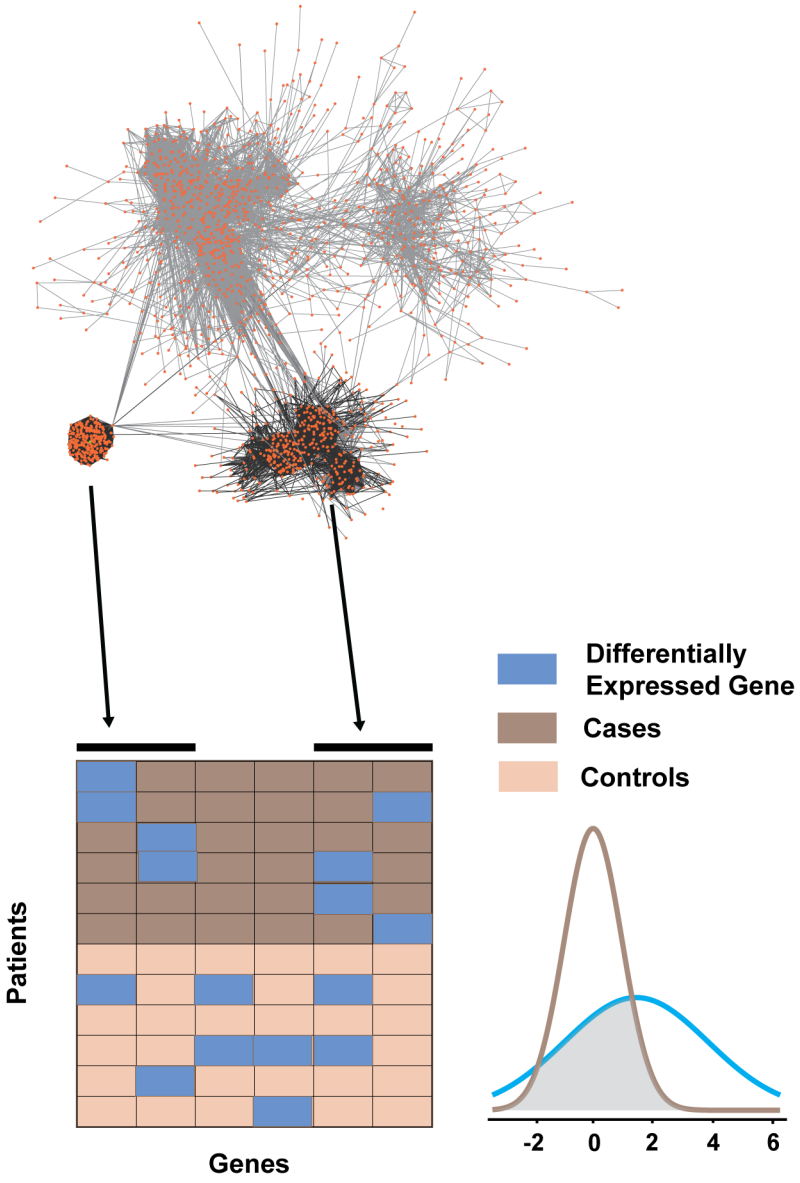


**Figure 2**: **Simulation Framework.** From a real network, pathways are chosen as network communities. Genes in the pathway get differentially expressed in a random subset disease patients by drawing their expression value from a normal distribution with $\mu = 1.5$ and $\sigma = 2.5$. All other genes draw their expression from standard normal distribution. Noise is added to the data by differentially expressing a few genes among normal patients also.

systems. We can suggest several more complex models which may possibly add more biology to the framework, but not without sacrificing some of the flexibility to explore the data characteristics in terms of signal and noise. To a network-based classifier, expression data and interaction networks are the only source of signal. In the simulation framework, expression signal can be increased by a) varying the number of disease genes or b) adjusting  to make a normal distribution that has little overlap with a standard normal distribution. Network signal, on the other hand, can be decreased by a) adding or remove edges from the network or b) picking disease genes from pathways that exhibit poor topological connectedness. With these four dials, we can inject different amounts of signal and noise into the data to recognize conditions where network information will be necessary for accurate classification.

## 2.2 Datasets

To evaluate network-classifiers, we chose five different expression studies. Among the five datasets, Wang (GSE 2034) [22], Bos (GSE 12276) [1] and Ivshina (GSE 4922) [12] datasets were downloaded from NCBIs Gene Expression Omnibus [7] and processed with RMA from affy R package [9]. van de Vijver dataset [21], along with its class labels, are the ones used in Dutkowski et al [6]. In both Wang and van de Vijver datasets, classification was done to differentiate metastatic versus non-metastatic patients. Ivshina et al. studied the ability of expressed profiles to identify different tumor grades. Since it was reported that grades G2 and G3 were difficult to classify using traditional approaches, we used them evaluate the power of network-based methods. Finally, Hernandez-Lobato et al. [11] showed the superiority of their network-based classification algorithm with expression data from Bos et al. We used the same dataset with patients split into two classes those with metastasis free survival of time less than 21 months and greater than 21 months. RNASeq data for breast cancer was downloaded from The Cancer Genome Atlas with the status of her2 as the class labels. Both for classification of real data as well as simulation, protein-protein interaction network downloaded

from STRING database v9.0 [20] was used after filtering for top 10% of edges based on the interaction score. In all the classification runs, only the genes present in both the interaction network and expression data were used.

## 2.3 Network-based classification algorithms

### 2.3.1 GraphSVM

It is well known that genes that interact have similar expression profiles. Going by this observation, GraphSVM reduces the noise in the expression of a gene by retaining only the signal components that are common among its neighbors and attenuating all other components. In other words, high frequency components, that are likely to be noise, are detected using the expression of its interacting partners and filtered out using techniques from spectral graph theory and discrete Fourier transform. Let $V$ be the set of vertices in the interaction network and let $L$ be the graph Laplacian of the network. Let $0 = \lambda_1 \leq ... \leq \lambda_n$ be the eigenvalues of $L$ and $e_1, e_2, ..., e_n$ be its eigenvectors. Since the eigen-basis of $L$ forms the Fourier basis, the discrete Fourier transform $\hat{f} \in \mathbf{R}^n$ of the expression data $f$ is defined as

$$\hat{f}_i = \sum_{u \in V} e_i(u) f(u), \quad i = 1, 2, ..., n; \tag{2.1}$$

Generally, the eigenvectors corresponding to larger eigenvalues tend to have higher variance on the graph, and are likely to be noise. Fourier transforms corresponding to each eigenvector is computed as a sum of the transforms on each node so as to capture the variance from the entire graph. Therefore, when recovering the expression matrix from the Fourier transform, an exponential decay function based on the eigenvalues is used to strongly attenuate high frequency components alone.

$$\forall f \in R^V, \quad S_\phi(f) = \sum_{i=1}^{n} \hat{f} \phi(\lambda_i) e_i \tag{2.2}$$

And,

$$\phi(\lambda_i) = exp(-\beta \lambda_i) \tag{2.3}$$

The Euclidian distance between two expression profiles based on $S_\phi$ can be expressed as inner products, which on simplification yields a positive semi-definite kernel function. This kernel is used in the support vector machine framework to construct an optimal-margin hyper-plane separating the expression vectors belong to the two classes.

In our experiments, R language [17] implementation of this algorithm provided in pathClass package [14] was used. However, the recursive feature elimination step added to the algorithm in pathClass was turned off to use the original method proposed in [18].

### 2.3.2 Network propagated SVM (NwpropSVM)

Network propagation is the idea of averaging the expression of genes over a connected network region based on the mechanics of water flowing through pipes. Assume that the edges of a network are pipes of equal capacity. Nodes act as tanks where water is pumped to as well as drained from. Water is pumped at constant rate through each node at quantity proportional to the expression of the gene. At every node, a constant amount of the stored water is lost at every time interval. Assuming at every time interval, water flows from one node to its neighbors, under this setup, a steady state will be reached at some time point where the change in water quantity at every node will be negligible. This quantity becomes the final expression level of that node. The new expression matrix based on the propagated expression levels is classified using a regular linear kernel support vector machine. We used the LibSVM [2] implementation through the R language interface provided by e1071 package [5]. This method is named as NwpropSVM in this thesis.

Intuitively, if all the genes in a pathway get over-expressed but in different subsets of disease patients, network propagation will result in a uniformly increased expression of all pathway genes in all disease patients. Such a prominent signal can be very easily learnt by classifier such as SVM, leading to improved sensitivity, specificity and accuracy of classification.

Mathematically, network propagation is very simple to compute. In order to normalize for the node degrees, adjacency matrix $A$ of the network is modified

as follows:

$$M = D^{(-0.5)}AD^{(-0.5)} \tag{2.4}$$

where $D$ is a diagonal matrix with the node degrees being the diagonal elements.

With the normalized adjacency matrix, iteratively propagate the expression of each gene as follows:

$$F_{i+1} = (\alpha \cdot F_i \cdot M) + (1 - \alpha)E; \quad F_0 = E \tag{2.5}$$

where $F_{i+1}$ denotes the expression matrix obtained after $i + 1$ rounds of network propagation, is the original expression matrix and is a scalar that controls the fraction of network information and original expression data that gets infused into the propagated expression matrix. This equation has a simple closed-form fixed point that is obtained by letting $F_{i+1} = F_i$ in Equation 2.5:

$$F = (1 - \alpha)(1 - \alpha M)^{-1}E \tag{2.6}$$

Even though the idea of network propagation had been proposed elsewhere in the literature, this is the first instance of its use in classification of gene expression data.

All the results reported in this work were obtained with a 5-fold cross validation that was repeated 100 times to compute the Area Under ROC Curve.

# Chapter 3

# Results

## 3.1 Better data is necessary to improve classification

The premise behind network-based classification is that gene expression is regulated by the interactions among genes and gene products. Therefore, in theory, classifiers that know these interactions can better estimate the impact of each genes expression on a phenotype based on the impact of its interacting partners, leading to improved classification. However, this theory did not hold in practice when we evaluated the performance of network-based classifiers on five different expression datasets (Figure 1D) NwpropSVM remains insensitive to network information, whether it is the real or a permuted network, showing no substantial increase in AUC over SVM; GraphSVM, on the other hand, is not even able to match the performance of SVM. To examine whether this trend is due to faulty algorithms, we turned to simulations where a gold-standard interaction network is available for classification. With a synthetic dataset generated as described in the Materials and Methods chapter, classification performance of GraphSVM and NwpropSVM is shown in Figure 3A. Both the classifiers show significant improvement in AUC because of the network, suggesting that network-constrained classifiers are useful, but with high quality datasets.

Parameters of the simulation are chosen such that a boost in performance

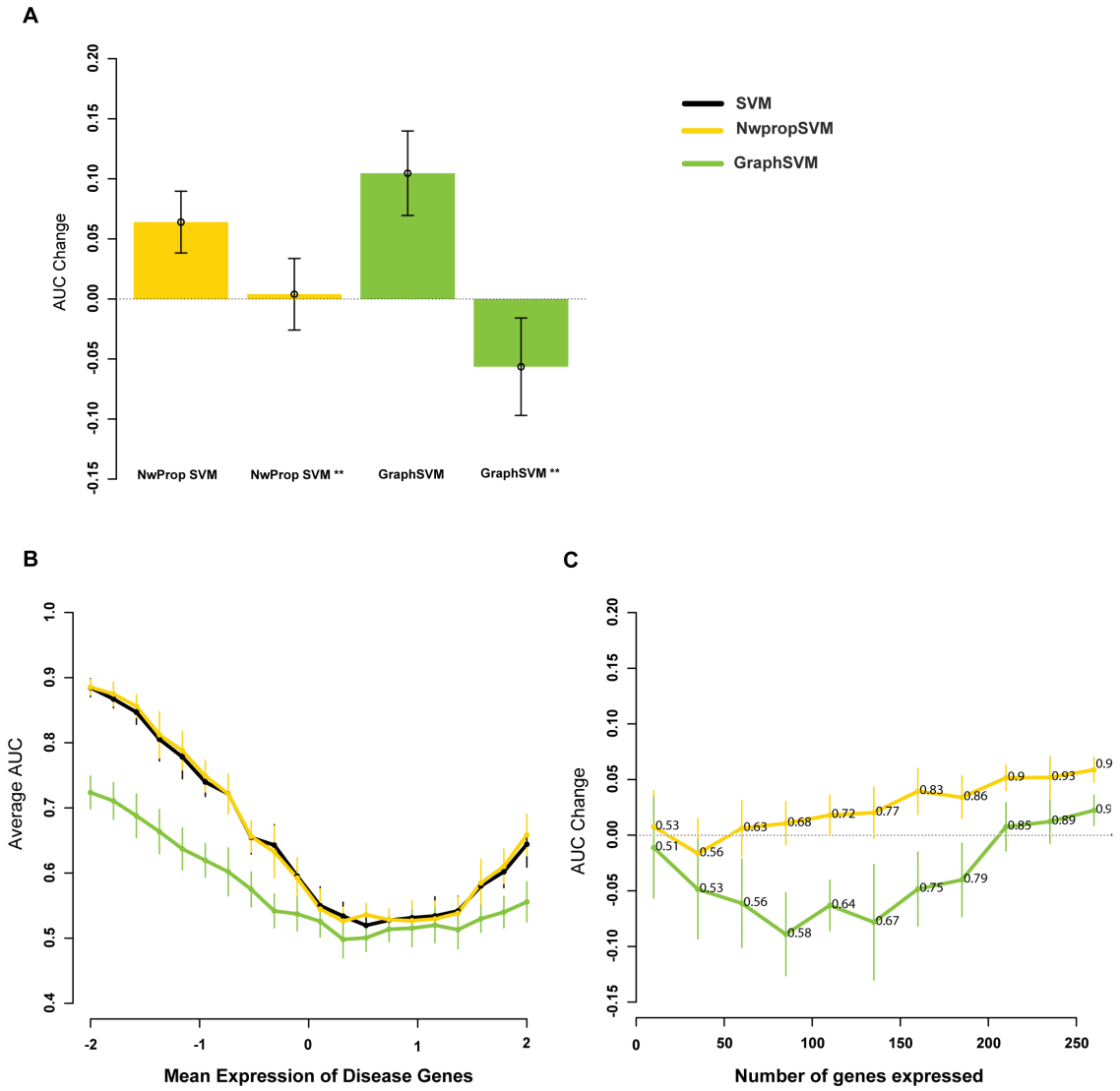


**Figure 3**: **Classification performance on simulated data.** With a gold-standard network used to simulate the expression data, network-based classifiers improve accuracy (Panel A). Simulation parameters were chosen such that signal from the generated expression data would be weak and that a network is necessary to realize substantial classification accuracy. Effect of expression signal is examined in Panels B and C where pathway signal is silenced by selecting disease genes at random. To minimize expression signal, simulation parameters were chosen to be the values in Panel B and C where both network-based methods yielded low AUC. In Panel C, the numbers on the line graph are the actual AUC values obtained by each classifier. They are showed to illustrate how increasing the number of expressed genes could easily benefit network classifiers despite not having any pathway signal.

on the synthetic dataset (Figure 3A) is only because of the network information and not due to the inherent signal in the expression profile. To this end, we simulated a new expression data where the signal from network was silenced by randomly selecting disease genes from the entire gene set instead of selecting from pathways. Signal from expression data was progressively increased to measure its independent effect on classification. Figure 3B shows classification performance as a result of varying the mean of expression of disease genes keeping standard deviation constant. Because all other genes draw their expression from standard normal distribution, high overlap of the two distributions and therefore low AUCs are expected around the mean value of zero. Even though neither network-constrained classifier gets undue boost in AUC in high signal regions, a mean value of 1.5 was chosen for the expression of disease genes for all the simulations results presented here. The idea is to keep the expression signal low such that even small improvements lent by the network would be apparent. Another source of signal in the expression data is the number of disease genes that show differential expression. Here again, disease genes were randomly selected and Figure 3C records the gain/loss in AUC of network-constrained methods over SVM. When more than hundred disease genes show differential expression, NwpropSVM shows an unjustified improvement in AUC over SVM, which could be simply an artifact of the network used. Therefore, throughout this work, a pathway having about fifty genes was picked for differential expression in simulation. Knowing the properties of simulated data, result from Fig 3A adds more confidence on the power of network-based classifiers, recapitulating the need for better data.

## 3.2 Noisy networks may or may not affect classification performance

Large-scale interaction networks are often a conglomeration of direct (physical) interactions and indirect (functional) interactions derived from high-throughput experiments, co-citations, computational predictions, and orthologous interactions found in other species. Such networks, are prone to noise in terms of false positive

and false negative interactions  interactions occurring within the cell but not captured by the network are false positives while those that do not occur in nature but present in the network are false negatives. When network-based classifiers prioritize feature selection based on genes connected in the network, false positives would cause genes to be missed from the selection while false negatives could lead classifiers to pick noisy gene combinations as predictive ones. Both forms of noise, therefore, are detrimental to classification performance.

Having already established that high quality networks would improve classification, we next chose to investigate how reducing the network quality would affect performance. Through simulation, an expression data is generated using pathway information from a network and classified using the same network but with noise added to it. False positives are added by removing edges and false negatives by adding edges to the network. While non-network classifiers will not be affected, network-constrained counterparts are susceptible to making wrong predictions based on noise. Figure 4 displays performance of SVM, NwPropSVM and GraphSVM as a function of noise in the network, in form of a heatmap. Darker shades indicate lower AUC and lighter shades denote higher AUCs. With the heatmap of SVM (Figure 4A) as the baseline, it can be observed that NwpropSVM (Figure 4B) performs no worse than SVM. GraphSVM (Figure 4C), on the other hand, exhibits a gradual decrease in AUC when more noise is added to the network.

In addition to measuring the impact of network noise, results from Figure 4 highlight a more important and often overlooked property of network-based classifiers  knowledge about the classification algorithm is necessary to summarize its effect on data. Performance of general purpose classifiers such as SVM is known to degrade with increase noise in the data. Since network-based classifiers are frequently an extension of the general purpose ones, common wisdom is that their performance would also degrade with increase in network noise. However, Figure 4 illustrates that even among two classifiers built on top of the same general purpose classification framework (SVM), one could be completely insensitive to noise and the other could be very sensitive. This observation is very important to solve the

puzzle that many network-constrained classifiers did not improve in performance compared to their unconstrained counterparts  while high quality networks aid in classification, noisy ones may or may not hurt performance depending on the specifics of the algorithm. In this context, we can offer a further interpration of the results in Figure 1D. If the STRING network was noisy, GraphSVM would exhibit poor classification while NwpropSVM would neither improve nor reduce performance when compared to SVM. This insight is also consistent with the result from Figure 1D that even on a permuted network which has no information, NwpropSVM would perform just as well as SVM.

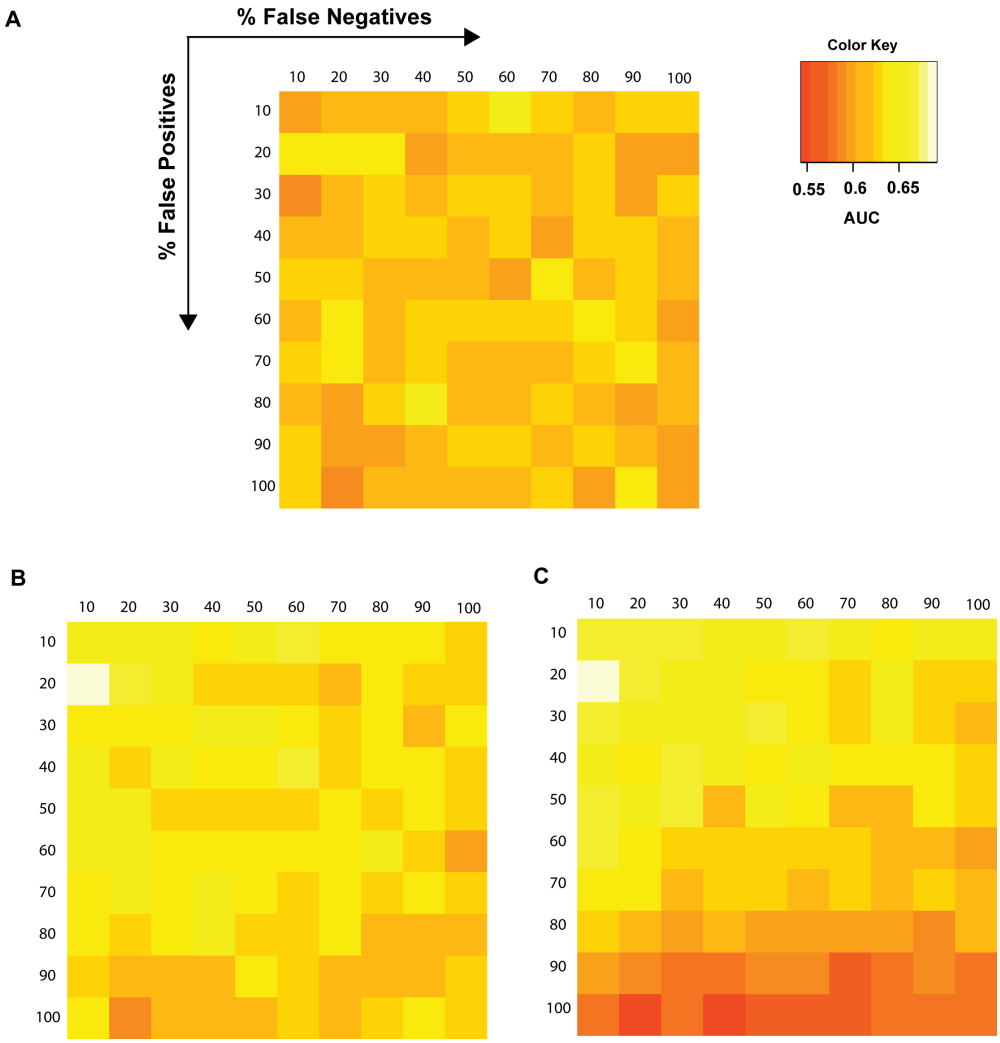


**Figure 4**: **Effect of network noise on classification performance.** Exploring all possible combinations of network noise by simultaneously adding False Positive and False Negative edges to the original network, a heatmap of AUCs for SVM (Panel A), NwpropSVM (Panel B) and GraphSVM (Panel C) is shown. Even though both network-based methods are extensions to SVM, they react in completely opposite ways to network noise. NwpropSVM is sensitive to network noise, but doesnt do any worse than regular SVM. GraphSVM, on the other hand, produces almost no useful classification with high amounts of netwprork noise.

# Chapter 4

# Conclusion and Future Work

Adding prior knowledge to aid molecular classification is an exciting and promising idea. Much success of these integrated classifiers is because of their ability to identify disease biomarkers as not just single genes but a combination of interacting genes. Network biomarkers tend to be stable across cohorts and therefore form excellent targets for therapeutic intervention. However, the main issue with many classifiers is that adding network information does not seem to improve sensitivity and specificity of the classification. In this paper, we investigated the reasons for this problem by simulating gene expression data using a known interaction network. This allows us to evaluate classifiers on this synthetic expression data using a gold-standard network to understand their behavior under ideal conditions. With simulation, it is possible to vary the proportions of signal and noise in the data to capture all conditions from ideal to imperfect.

Under ideal conditions, all classifiers that are considered here realize a boost in AUC when using network information. However, when adding noise to the network, NwpropSVM perform less than ideal but no worse than SVM. On the contrary, performance of GraphSVM degrades gradually with increase in amount of noise in the network and reaching AUCs close to 0.5 when there is no more signal. Reasoning the results from Figure 1D using this trend, it is conceivable that the STRING network used for classification is not adding any information. The actual network is not necessarily uninformative, but for the five expression studies

considered in this work, it seems to be a poor prior. Four (Wang et al.[22], van de Vijver et al.[21], Bos et al.[1], Ivshina et al.[12]) of the five expression data had been previous used for network-based classification that showed an improvement in performance. Therefore, the expression of genes in the data follows a particular pattern but the pattern is not properly captured by the STRING network.

Network biomarkers promise to improve the accuracy of classification as well as reproducibility across datasets. In this work, we have shown that even though network-based classifiers do not improve accuracy on real datasets, they would work as expected when the quality of datasets is improved. However, we demonstrated the results through AUCs derived from cross-validation on a single dataset. In the future, we hope to examine how these methods perform when trained on one dataset and tested on another. The current simulation framework has to be extended to generate two datasets that are based on the same prior knowledge but differ in the process of generation. This would be essential to capture the variances between two expression studies on different cohorts and different measurement platforms.

# Bibliography


[1] Bos, P. D., Zhang, X. H., Nadal, C., Shu, W., Gomis, R. R., Nguyen, D. X., Minn, A. J., van de Vijver, M. J., Gerald, W. L., Foekens, J. A., and Massague, J. Genes that mediate breast cancer metastasis to the brain. *Nature 459*, 7249 (2009), 1005–9.

[2] Chang, C.-C., and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

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

[4] Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20* (1995), 273–297. 10.1007/BF00994018.

[5] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., , and Weingessel, A. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2011. R package version 1.6.

[6] Dutkowski, J., and Ideker, T. Protein networks as logic functions in development and cancer. *PLoS Comput Biol 7*, 9 (2011), e1002180.

[7] Edgar, R., Domrachev, M., and Lash, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res 30*, 1 (2002), 207–10.

[8] Ein-Dor, L., Zuk, O., and Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences 103*, 15 (2006), 5923–5928.

[9] Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics 20*, 3 (2004), 307–15.

[10] Guillemot, V., Tenenhaus, A., Le Brusquet, L., and Frouin, V. Graph constrained discriminant analysis: A new method for the integration of a graph into a classification process. *PLoS ONE 6*, 10 (10 2011), e26146.

[11] Hernandez-Lobato, J. M., Hernandez-Lobato, D., and Suarez, A. Network-based sparse bayesian classification. *Pattern Recognition 44*, 4 (2011), 886–900.

[12] Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J. E., Liu, E. T., Bergh, J., Kuznetsov, V. A., and Miller, L. D. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res 66*, 21 (2006), 10292–301.

[13] Johannes, M., Brase, J. C., Frhlich, H., Gade, S., Gehrmann, M., Flth, M., Sltmann, H., and Beibarth, T. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics 26*, 17 (2010), 2136–2144.

[14] Johannes, M., Frohlich, H., Sultmann, H., and Beissbarth, T. pathclass: an r-package for integration of pathway knowledge into support vector machines for biomarker discovery. *Bioinformatics 27*, 10 (2011), 1442–3.

[15] Lavi, O., Dror, G., and Shamir, R. Network-induced classication kernels for gene expression prole analysis. http://acgt.cs.tau.ac.il/papers/NICKs.pdf.

[16] Li, C., and Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics 24*, 9 (2008), 1175–1182.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[18] Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J. P. Classification of microarray data using gene networks. *BMC Bioinformatics 8* (2007), 35.

[19] Ruan, J., and Zhang, W. Identifying network communities with a high resolution. *Phys. Rev. E 77* (Jan 2008), 016104.

[20] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res 39*, Database issue (2011), D561–8.

[21] van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas,

A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med 347*, 25 (2002), 1999–2009.

[22] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet 365*, 9460 (2005), 671–9.

[23] Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knsel, T., Rmmele, P., Jahnke, B., Hentrich, V., Rckert, F., Niedergethmann, M., Weichert, W., Bahra, M., Schlitt, H. J., Settmacher, U., Friess, H., Bchler, M., Saeger, H.-D., Schroeder, M., Pilarsky, C., and Grtzmann, R. Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol 8*, 5 (05 2012), e1002511.

[24] Zhu, Y., Shen, X., and Pan, W. Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics* (2009), 1–21.