Module-Based Biomarker Discovery in Breast Cancer

Yuji Zhang[1,2], Jason J. Xuan[1], Robert Clarke[2], Habtom W. Ressom[2,*]

[1]*Dept. of Electrical & Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA*
[2]*Lombardi Comprehensive Cancer Center, Georgetown University medical Center, Washington, DC*
[*]*hwr@georgetown.edu*

Abstract

The availability of genome-wide biological network data opens up new possibilities to discover novel biomarkers and elucidate cancer-related complex mechanisms at network level. In this paper, we propose a novel module-based feature selection framework, which integrates biological network information and gene expression data to identify biomarkers, not as individual genes but as functional modules. Also, a large-scale analysis of ensemble feature selection concept is presented. The method allows combining features selected from multiple runs with various data subsampling to increase the reliability and classification accuracy of the final set of selected features. The results from four breast cancer studies demonstrate that the identified module biomarkers have the following important features: i) achieve higher prediction accuracy in independent validation datasets; ii) are more reproducible than individual gene biomarkers; iii) improve the biological interpretability of results; and iv) are enriched in cancer-related "disease drivers".

1. Introduction

In the last few decades, high-throughput genomic and proteomic techniques have generated a large number of diagnostic, prognostic and predictive molecular signatures related to many diseases [1]. Traditional gene biomarkers are typically selected by scoring individual genes for how well their expression patterns discriminate between different subclasses of disease or between cases and controls. However, there are several disadvantages of these approaches including the following: (1) Lack of adequate biological interpretation, because the genes selected by traditional biomarker discovery methods are mainly "downstream" reflectors of the perturbations defining clinical outcomes through the complex interplay of biological networks. Thus, they may not directly account for the activity, perturbations or roles that disease-related cellular networks show. (2) Oversimplified assumption of gene independence, i.e., traditional biomarker discovery approaches assume gene independence. (3) Low reproducibility/reliability, i.e., biomarker sets identified from different labs share very few genes in common. (4) Inadequate focus on genes that are "disease drivers" such as oncogenes and tumor suppressors whose mutations result in a detrimental change of function that leads to cancer.

The above limitations of traditional biomarker discovery approaches have received great attention by the community of cancer research [2]. We argue that the fundamental reason for these limitations is that these traditional biomarker identification methods lead to genes whose roles are mostly "passengers" rather than "drivers" of the phenotypic differences between sample groups (e.g., poor versus good outcomes). Regulatory networks often act as amplification cascade, where highly differentially expressed genes tend to be further downstream from the somatic or inherited determinants of the clinical outcomes. Since the regulatory networks comprise the complex interactions of multiple potential casual factors and sources of biological noise [3], these downstream genes are more prone to be most unstable across and within samples. On the other hand, oncogenes and tumor suppressors are generally not the most differentially expressed genes although they may show an outlier behavior in some samples [4]. The biomarkers enriched in these disease drivers may represent upstream regulators with potential causal roles in the determination of differential phenotypes, which will improve the reliability and reproducibility of the prediction model in unknown samples. For instance, Lim et al. succeeded in detecting candidate biomarkers by identifying "upstream regulators" causally related to the phenotypic differences [5]. The inferred sets of "master regulators" were shown to be more powerful and robust than the signatures proposed by original investigations based on standard gene-based analysis. Such studies imply that systems approaches to biomarker discovery in a biological network context would identify biomarkers more indicative of phenotypic changes.

The availability of biological network data enables new opportunities for elucidating modules involved in major diseases and pathologies. Several approaches have been demonstrated to extract the relevant functional modules based on coherent expression patterns of their genes [6, 7]. However, these biological interaction networks have been typically analyzed separately in previous studies. Such approaches tend to hide the full complexity of the cellular circuitry since many processes involve combinations of different types of interactions.

In this paper, we propose a module-based systems biology approach to identifying module biomarkers of diseases by integrating patient gene expression profiles and different types of biological network data, including protein-protein interaction network, protein-DNA interaction network, and signaling pathway network. The biomarkers here are not encoded as individual genes or proteins, but as modules of interacting proteins within a large-scale human interaction network. Analysis of four breast cancer cohorts shows that this method has several advantages over previous analyses of differential expression. First, the resulting module biomarkers provide models of the molecular mechanisms underlying disease mechanisms. Second, module-based classification achieves higher accuracy in prediction, which is ascertained by selecting biomarkers from a training set and evaluating them on an independent validation set. Third, the identified module biomarkers are more reproducible in different experiments than individual biomarker genes selected without network information. Also, our approach provides the capability to detect genes with known disease mutations that are typically not detected through gene-based differential expression analysis. These genes are referred to as "disease drivers" that are causally responsible for the determinations of differential phenotypes.

2. Materials and methods

In this section, we describe the data used in this work, followed by a description of our proposed module-based biomarker discovery approach that integrates gene expression profiles and biological network information. Fig. 1 illustrates the steps of the proposed approach.

2.1 Datasets

Gene expression data: we obtained three mRNA expression datasets from three breast cancer studies [8-10] and one in house dataset. We divided these datasets into two groups: (1) prognosis group and (2) endocrine treatment prediction group. The prognosis group includes the van de Vijer and the Wang datasets that consist of patients with either poor or good outcomes. Poor outcome is defined as patients with time of metastasis within five years of surgery, and good outcome as those with time of metastasis greater than or equal to five years after surgery. The endocrine treatment prediction group includes the Loi and our in house datasets consisting of patients with either early recurrence or non-recurrence. Early recurrence is defined as patients with recurrence within three years of endocrine treatment, and non-recurrence refers to those with time of recurrence greater than fifteen years after endocrine treatment. Table 1 presents the statistics of each dataset and the corresponding microarray platform. Since the four studies were performed on different microarray platforms, we restrict our analysis to the common genes present in all datasets. For simplicity, we used the terms "gene" and "protein" interchangeably in this work.

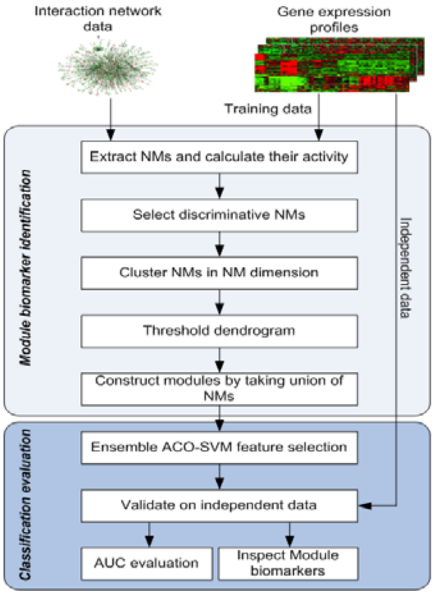
Interaction network data
Gene expression profiles
Training data
Module biomarker identification
Extract NMs and calculate their activity
Select discriminative NMs
Cluster NMs in NM dimension
Threshold dendrogram
Construct modules by taking union of NMs
Independent data
Classification evaluation
Ensemble ACO-SVM feature selection
Validate on independent data
AUC evaluation
Inspect Module biomarkers

Fig. 1. Overview of module-based biomarker identification.

We normalized the expression of each gene across all samples in every dataset separately. For the dataset generated by Agilent platform, we used log ratio (base 2) between the measured and control samples. For datasets generated by Affymetrix chips, we used log (base 2) to transform the original expression values of each gene in each array. For both types of datasets, we normalize the log-space gene expression values by

$$g_{ij} \rightarrow \log_2(g_{ij}) - \log_2(\overline{g_i}) = \log_2(\frac{g_{ij}}{g_i}) \qquad (1)$$

where g_{ij} is the intensity of gene i on a particular sample j, and $\overline{g_i}$ is the mean intensity of gene g_i over all samples. This normalization mimics a two channel microarray where the reference channel is a pool of all samples under consideration [11].

Table 1. Four datasets used for evaluation.

Name	Microarray platform	Number of samples
van de Vijver dataset	Agilent oligonucleotide Hu25K	Poor outcome: 78 samples Good outcome: 217 samples
Wang dataset	Affymetrix HG-U133a	Poor outcome: 106 samples Good outcome: 180 samples
Loi dataset	Affymetrix HG-U133	Early recurrence: 12 samples Non recurrence: 12 samples
In house dataset	Affymetrix HG-U133	Early recurrence: 24 samples Non recurrence: 40 samples

Biological network data: Protein-protein interaction and protein-DNA interaction data were extracted as previously described [12]. Signaling network data were extracted from the following three sources: i) the most comprehensive signaling pathway database, BioCarta (http://www.biocarta.com/); ii) a literature-mined signaling network [13]; and iii) 10 manually curated signaling pathways from the Cancer Cell Map (http://cancer.cellmap.org/cellmap). In this work, we built a common interaction network containing 63,113 protein-protein interactions, 1789 protein-DNA interactions and 3,862 signaling interactions among 10650 common genes in all four datasets.

2.2 Module biomarker identification

To detect the modules in the collected biological network, we first extracted the significant network motifs (NMs) in the integrated cellular network as previously described [14]. NMs are statistically significant recurring structural patterns that are found more often in a real network than that would be expected in a random network with same network topologies. They are the smallest basic functional and evolutionarily conserved units in the biological network. Cancer-related genes have been shown to be more conserved compared to other genes along evolution [15, 16]. We assume that NMs in a biological network are enriched in "disease driver" genes which are more conserved than other downstream "passenger" genes. These NMs could form large aggregated modules that perform specific functions by forming collaborations among a large number of NMs. In this work, we focused on three-node NMs since larger size NMs (number of nodes > 3) are composed of three-node ones in most cases [17].

All the identified NMs were examined by calculating their activity scores via gene expression data. Each NM is considered as a subnetwork. We assume that in a subnetwork A, there are M genes with expression levels across N patient samples:

$$G_k = \{g_{ij} \mid i = 1,2,..., M, j = 1,2,..., N\} \qquad (2)$$

Given a particular gene i, the expression values g_{ij} are normalized to z-transformed scores z_{ij} so that the z score vector z_i has mean $\mu = 0$ and standard deviation $\sigma = 1$ over all samples j. The z score is defined by

$$z_{ij} = \frac{g_{ij} - \hat{\mu}_i}{\hat{\sigma}_i} \qquad (3)$$

where $\hat{\mu}_i$ is mean expression value of gene i across samples, and $\hat{\sigma}_i$ is standard deviation of expression value of gene i across samples.

Let z represent the corresponding vector of class labels (e.g., tumor metastatic or non-metastatic). The discriminative score of gene i is defined as the mutual information $MI_i(x;y)$ between the expression levels of gene i and sample labels c:

$$MI_i(x;y) = \sum_{x \in z_i} \sum_{y \in c} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad (4)$$

where x is the discretized value of z_i, and y is the sample lables, $p(x,y)$ is the joint probability density function of z_i and c, and $p(x)$ and $p(y)$ are the marginal pdf's of z_i and c. A histogram technique is applied to transform the continuous gene expression values to discrete ones for the calculation of the mutual information [18].

The activity score of a subnetwork A was calculated by combining the transformed z scores of its individual genes. The individual z_{ij} of each member gene in one subnetwork were combined into z_{A_j} by

$$z_{A\_j} = \frac{1}{\sqrt{\sum_{i=1}^{M} w_i^2}} \sum_{i=1}^{M} w_i z_{ij} \qquad (5)$$

where w_i denotes the weight that is defined as

$$w_i = \frac{MI_i(x;y)}{\sum_{i=1}^{M} MI_i(x;y)} \qquad (6)$$

The weighted z score is intended to emphasize the hub genes which are surrounded by many highly discriminative genes although they are not highly differently expressed themselves.

The discriminative score of subnetwork A is calculated similarly as defined in Eq. (4):

$$MI_A(x;y) = \sum_{x \in z_A} \sum_{y \in c} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad (7)$$

where x is the discretized value of z_A, and y is the sample labels.

We performed two permutation tests to assess the significance of the identified NMs. For the first test, we tested whether the mutual information with the disease class is stronger than that obtained with random assignments of classes to patients [19]. For the random model, we permuted the sample labels for 100000 trials, yielding a null distribution of mutual information scores for each trial, and the real score of each NM was indexed on this null distribution. For the second test, we tested if the mutual information with network interactions was stronger than that obtained with random assignments of gene expression vectors to individual genes. The mutual information for each NM was calculated over 100000 random trials in which the expression vectors of individual genes were permuted over the network. The score of each NM was indexed on the "global" null distribution of all random NM activity scores. In this work, significant NMs were selected with both *P* values less than 0.0001.

The NMs that passed the significance tests were clustered in the NM dimension using the hierarchical clustering method. This resulted in a tree in which each internal leaf node is associated with a vector representing the average of all of the NM vectors at its decent leaves. We annotated each interior node with the Pearson correlation between the vectors associated with its two children in the hierarchy. We defined as NM cluster in which each interior node whose Pearson correlation differed by more than 0.05 from the Pearson correlation of its parent node in the hierarchy. The module was then formed by taking the union of the clustered NMs.

2.3 Ensemble classification evaluation

In order to select robust module biomarkers for classification of unknown patient samples, we applied an ensemble feature selection technique to select module subsets in training dataset and validate their discriminative power in an independent validation dataset. Similar to ensemble learning for classification, ensemble feature selection techniques use a two-step procedure: i) a number of different feature selectors are created; ii) the outputs of these component feature selectors are aggregated to generate the final ensemble results. We focused on the analysis of ensemble feature selection techniques using ant colony optimization – support vector machine (ACO-SVM) feature selection approach we previously developed [20]. The ACO-SVM approach was used to select the best features in terms of their ability to distinguish between two patient phenotypes in a validation dataset which were not involved in the feature selection step.

To obtain a robust module biomarker set in one dataset, we generated slight variations of the original dataset, and aggregated the features selected by ACO-SVM from these variant samples. The rationale behind this is that for a stable biomarker set, training datasets with small change should generate biomarker sets with high similarities. The biomarkers with high frequencies in these biomarker sets are presumed to be most relevant to sample distinction and used to predict the class membership of independent samples. A subsampling approach was proposed to generate the training datasets with slight variations: a large number (e.g., 500) of datasets are generated by stratified subsampling the original dataset without replacement. As gene expression datasets generally contain only tens of samples, we generated subsamplings containing 90% of the samples of the original dataset, and the remaining 10% of the samples were used as internal validation dataset to estimate the performance of a classifier, called *within-dataset validation*. Since we considered typically 500 independent partitions in 90% training and 10% validation, we reduced the risk of overoptimistic results of traditional cross-validation experiments on small sample domains [21].

The biomarker sets generated from 500 subsampling datasets using the ACO-SVM approach were then evaluated through a frequency plot, where we computed the frequency with which modules were selected was then analyzed. The most frequently selected set of modules was then validated by using it to classify an independent validation dataset. This approach is referred to as *cross-dataset validation*.

The double-validation procedure stated above was designed to provide an unbiased evaluation of the generalization error in independent dataset. Since both prognostic and treatment outcome prediction groups contain two datasets, we evaluated the classification performance of the module biomarker set generated from one dataset on the other dataset in the same group, or vice versa.

3. Results

We report here the experimental evaluations of our methods to search for module biomarkers with discriminative power between different subgroups of breast cancer patients in a biological network context. Four breast cancer datasets were used to identify potential biomarkers.

The collected biological network involved 72,562 three-node NMs detected using FANMOD tool [22]. Totally, 1017, 752, 696 and 908 NMs were identified in the four breast cancer datasets (van de Vijer, Wang, Loi, and in house datasets, respectively). This is based on two permutation tests for statistical significance consisting of 581, 707, 793, and 886 genes, respectively. Using hierarchical clustering analysis, 162, 313, 270 and 343 module markers were constructed as candidate module biomarkers of the four datasets, respectively. Each module may be viewed as a putative marker for breast cancer. The modules are not based on individual detected genes, but rather on the aggregate behavior of genes connected in a functional module. This approach is indeed a departure from conventional gene-based analysis, which does not provide biological insight into the identified markers.

Table 2. BCGs in module markers of four datasets.

BCG	van de Vijer	Wang	Loi	In house
Differentially expressed (P<0.05)	10	15	16	13
Not differentially expressed	61	64	62	88
Total	71	79	78	131

We investigated whether the proposed module-based analysis can implicate upstream disease driver genes with relative low discriminative potential (e.g., those with larger *P* value in two-tailed t-test). Such proteins can arise within a significant module if they are essential for maintaining its integrity. Moreover, these disease driver genes are mostly in the upstream of the gene regulatory cascade, regulating their downstream genes to be differentially expressed under different disease status. Detecting modules containing these disease driver genes is expected to improve the reliability and robustness of these module biomarkers across different datasets. To evaluate the power of a module-based method to identify disease driver genes, we assembled a list consisting of 711 breast cancer genes (BCGs) extracted from the Online Mendelian Inheritance in Man (OMIM) database. The genes in the module markers identified from four datasets are more enriched with these BCGs than the ones from a conventional gene expression based analysis without network information. In particular, we found that 69 out of 162, 123 out of 313, 120 out of 270, and 136 out of 343 module markers contained at least one known BCG. We observed that 31, 26, 41 and 44 module biomarkers contained two or more known BCGs, respectively. Most of these BCGs were not significantly differently expressed (Table 2). Disease genes that can be only detected by the proposed approach include BRCA1, ESR1, TP53, etc.

We also examined the agreement between module markers identified from different cohorts of patients. The same classification process was also run for gene biomarkers selected by conventional methods. For comparison purpose, the top 581, 707, 793, and 886 discriminative genes in four datasets, respectively, were used as inputs to the classification process, which is the same number of genes covered by the module biomarkers for four datasets. As shown in Fig. 2, the module markers are more reproducible between datasets than individual marker genes selected without network information (e.g. t-test).

We tested the classification ability of the module biomarkers identified from four datasets by ACO-SVM. To use module information for classification, the weighted *z* score of module biomarkers were used as input feature values to a classifier based on SVM. An ensemble ACO-SVM approach was used to select the optimal features based on Area Under the ROC Curve (AUC) scores in a double-validation procedure, as described in Methods section. We used a baseline ACO-SVM approach for comparison purpose. To perform ensemble feature selection for gene biomarkers, the *z* score of candidate gene biomarkers were used as input feature values to a classifier based on SVM. The AUC scores of the second independent validation dataset by the classifier built from both module and gene biomarkers selected from the first dataset are depicted in Fig. 3, which shows that the module biomarkers outperform the gene biomarkers in all four experiments. This implies that the module biomarkers are more robust across different datasets generated on different platforms.

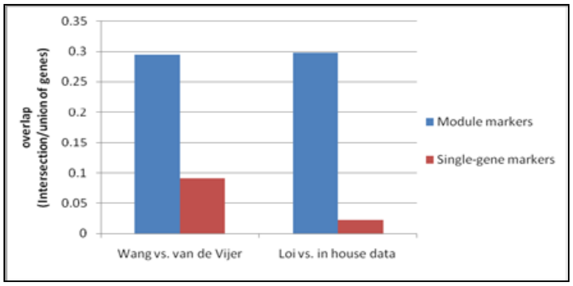

Fig. 2. Agreement in markers selected from one dataset versus those selected from the other dataset in the same clinical group.

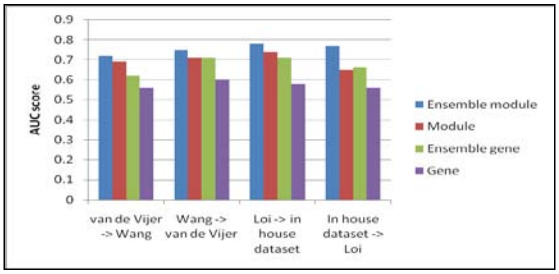

Fig. 3. AUC classification performance of modules, genes with ensemble feature selection strategy, and without the ensemble strategy.

4. Discussions

In this paper, we introduce a module-based feature selection framework to identify module biomarkers that lead to high reproducibility and classification accuracy. This is accomplished by a novel hybrid feature selection approach that identifies groups of associated genes by incorporating biological network information.

Several studies have been reported to integrate interaction network information and other biological data (e.g., gene expression data) for identification of genetic mediators of disease progression. However, only individual interaction layers, such as the transcriptional layer or the protein complex layer, were modeled by these methods [6, 7]. We propose an integrative approach for the identification of module-based biomarkers associated with the presentation of a specific tumor phenotype. In our approach, we chose to use a biological network containing protein-protein, protein-DNA and signaling pathway information. By adopting a genome-wide, mixed-interaction network, instead of the individual interaction layers of previous studies, we cover a far greater range of processes within the cell. This integration allows the method to capture several different mechanisms of action associated cancer progression and metastasis.

Compared to Lee et al. [6] and Chuang et al. [7], besides larger coverage of biological processes in our analysis, our approach utilizes an ensemble feature selection method to improve the classification accuracy and reliability of the module biomarkers. Both Chuang et al. and Lee et al. applied five-fold cross validation for one single dataset, which would generate overoptimistic results that do not adequately reproduce in independent datasets. In this work, a strict double validation strategy was used to estimate the classification performance. Such strategy leads to more reliable module markers than the gene-based approach. The relatively better overlaps of the module biomarkers derived from different datasets confirm that the module biomarkers are likely to be involved in cancer related mechanisms.

Acknowledgements

This research was supported in part by National Institutes of Health Grants (R21CA139246 and R01GM086746) and National Science Foundation Grant (IIS-0812246).

5. References

[1] R. Clarke, H. W. Ressom, et al., "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat Rev Cancer*, vol. 8, pp. 37-49, 2008.

[2] L. Ein-Dor, I. Kela, et al., "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics*, vol. 21, pp. 171-8, 2005.

[3] F. Azuaje, "What does systems biology mean for biomarker discovery?," *Expert Opinion on Medical Diagnostics*, vol. 4, pp. 1-10, 2010.

[4] S. A. Tomlins, D. R. Rhodes, et al., "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science*, vol. 310, pp. 644-8, 2005.

[5] W. K. Lim, E. Lyashenko, et al., "Master regulators used as breast cancer metastasis classifier," *Pac Symp Biocomput*, pp. 504-15, 2009.

[6] E. Lee, H. Y. Chuang, et al., "Inferring pathway activity toward precise disease classification," *PLoS Comput Biol*, vol. 4, pp. e1000217, 2008.

[7] H. Y. Chuang, E. Lee, et al., "Network-based classification of breast cancer metastasis," *Mol Syst Biol*, vol. 3, pp. 140, 2007.

[8] S. Loi, B. Haibe-Kains, et al., "Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen," *BMC Genomics*, vol. 9, pp. 239, 2008.

[9] M. J. van de Vijver, Y. D. He, et al., "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med*, vol. 347, pp. 1999-2009, 2002.

[10] Y. Wang, J. G. Klijn, et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, pp. 671-9, 2005.

[11] E. Segal, N. Friedman, et al., "A module map showing conditional activity of expression modules in cancer," *Nat Genet*, vol. 36, pp. 1090-8, 2004.

[12] Y. Zhang, J. Xuan, et al., "Network motif-based identification of breast cancer susceptibility genes," Conf Proc IEEE Eng Med Biol Soc. 2008, pp. 5696-5699, 2008.

[13] A. Ma'ayan, S. L. Jenkins, et al., "Formation of regulatory patterns during signal propagation in a Mammalian cellular network," *Science*, vol. 309, pp. 1078-83, 2005.

[14] R. Milo, S. Shen-Orr, et al., "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, pp. 824-7, 2002.

[15] A. Awan, H. Bari, et al., "Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network," *IET Syst Biol*, vol. 1, pp. 292-7, 2007.

[16] S. Narsing, Z. Jelsovsky, et al., "Genes that contribute to cancer fusion genes are large and evolutionarily conserved," *Cancer Genet Cytogenet*, vol. 191, pp. 78-84, 2009.

[17] E. Yeger-Lotem, S. Sattath, et al., "Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction," *Proc Natl Acad Sci U S A*, vol. 101, pp. 5934-9, 2004.

[18] G. D. Tourassi, E. D. Frederick, et al., "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Med Phys*, vol. 28, pp. 2394-402, 2001.

[19] L. Tian, S. A. Greenberg, et al., "Discovering statistically significant pathways in expression profiling studies," *Proc Natl Acad Sci U S A*, vol. 102, pp. 13544-9, 2005.

[20] H. W. Ressom, R. S. Varghese, et al., "Peak selection from MALDI-TOF mass spectra using ant colony optimization," *Bioinformatics*, vol. 23, pp. 619-26, 2007.

[21] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, pp. 374-80, 2004.

[22] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, pp. 1152-3, 2006.