Published in final edited form as:
Int J Data Min Bioinform. 2013 ; 7(3): 284–302.

Module-Based Breast Cancer Classification

Yuji Zhang[1], **Jianhua Xuan**[2], **Robert Clarke**[3], and **Habtom W. Ressom**[3,*]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, Minnesota, 55905, USA

[2]Department of Electrical & Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, Virginia, 22203, USA

[3]Lombardi Comprehensive Cancer Center, Georgetown University medical Center, Washington, District of Columbia, 20057, USA

Abstract

The reliability and reproducibility of gene biomarkers for classification of cancer patients has been challenged due to measurement noise and biological heterogeneity among patients. 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. Results from four breast cancer studies demonstrate that the identified module biomarkers i) achieve higher classification accuracy in independent validation datasets; ii) are more reproducible than individual gene markers; iii) improve the biological interpretability of results; and iv) are enriched in cancer "disease drivers".

Keywords

Cancer biomarkers; systems biology; feature selection; disease classification

1. Introduction

Over 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 [1–7]. Traditional biomarker discovery approaches for disease classification are typically selected by scoring individual genes for how well their expression pattern discriminate between different subclasses of disease or between cases and controls. However, there are several disadvantages of these approaches including the following:

Lack of adequate biological interpretation: 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. They may not directly account for the activity, perturbations or roles that disease-related cellular networks show [8].

Oversimplified assumption of gene independence: traditional biomarker discovery approaches make biological and statistical assumption of between gene independence, i.e., gene biomarkers are typically selected independently although proteins are well known to

*Corresponding author: Address: Suite 173, Building D, 4000 Reservoir Road NW, Washington DC 20057, Tel: (202) 687- 2283, Fax: (202) 687- 0227, hwr@georgetown.edu.

function coordinately within protein complexes, signaling pathways, and higher-order cellular processes. Thus, the resulting classifiers may contain marker genes with redundant information that may lead to decreased classification performance.

Low reproducibility/reliability: biomarker sets identified from different labs share very few genes in common. This is well illustrated by two prominent studies of survival prediction in breast cancer. van't Veer et al. [9] generated a list of 70 genes from 96 patient samples, which were subsequently tested successfully on a larger cohort of 295 patients [2]. Wang et al. [10] analyzed the gene expression profiles of 286 patients and reported a gene biomarker set of 76 genes. Each gene set was trained and tested within its own samples and achieved good prediction performance. However, the overlap of these two gene sets was very small: only three genes are in common. As a result, the predictive power of a classifier developed from one study could not be adequately reproduced when testing it on samples of another study, although both studies contain patients with similar phenotypes. Cellular heterogeneity within tissues and genetic heterogeneity across patients in complex diseases (e.g., breast cancer) may weaken the discriminative power of individual genes, even within a clinically homogeneous patient group [11].

Inadequate focus on genes that are "disease drivers": oncogenes and tumor suppressors are disease drivers whose mutations result in a detrimental change of function that leads to cancer. These genes are generally more conserved than other proteins, and tend not to be highly differentially expressed between different clinical groups of patients [12]. These genes would not be selected by traditional statistical ranking methods, such as TP53 and MYC. However, their expression patterns are more stable in patients of the same clinical subgroups and more robust across different studies. Search for biomarkers that may represent upstream regulators with potential causal roles in the determination of differential phenotypes may help us define more reliable and reproducible biomarker sets.

The above limitations of traditional biomarker discovery approaches have received great attention by the community of cancer research [11, 13–15]. 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 [16], 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 [17]. 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. Lim et al. succeeded in detecting candidate biomarkers by identifying "upstream regulators" causally related to the phenotypic differences [18]. In Lim et al., the transcriptional factors were determined if they caused the up/down-regulation of genes linked to poor outcome through patient samples in gene interaction networks inferred by ARACNe algorithm [19]. 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 large protein-protein interaction, protein-DNA interaction, and signal transduction pathway data enables new opportunities for elucidating modules involved in major diseases and pathologies [20]. Several approaches have been demonstrated to extract the relevant functional modules based on coherent expression patterns of their genes [21, 22]. However, these biological interaction networks have been typically analyzed separately in previous studies [21–24]. Such approaches tend to hide the full complexity of the cellular circuitry since many processes involve combinations of different types of interactions.

In this work, we propose a novel 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. Our experiments on four breast cancer cohorts show that the proposed 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 markers from a training set and evaluating them on an independent validation set. Third, the identified module biomarkers are likely to be more reproducible between different disease experiments than individual marker 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

We describe here the data and the methods we used in this study. The data consist of four breast cancer gene expression datasets and human interaction data collected from public databases. Our proposed module-based biomarker discovery approach integrates gene expression profiles and biological network information. Figure 1 illustrates the steps involved in this approach. In the following sections, we briefly describe the data and the analytical steps presented in Figure 1.

2.1 Datasets

Gene expression data: we obtained three mRNA expression datasets from three breast cancer studies [2, 10, 25] 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 all 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 number of patients in each dataset and the microarray platform used to generate gene expression data. Since the four studies were performed on different microarray platforms, we restrict our analysis to the common genes present in all datasets (all probesets were mapped to gene Entrez IDs). For simplicity, we used the terms “gene” and “protein” interchangeably in this work.

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

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

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 [26].

Biological network data: protein-protein interaction data were extracted from eight protein interaction databases [27–34] and two high-throughput yeast two-hybrid studies [35, 36]. Protein-DNA interaction data were extracted from the TRANSFAC database [37]. Signaling network data were extracted from the following three sources: i) manually curating the most comprehensive signaling pathway database, BioCarta (http://www.biocarta.com/); ii) a literature-mined signaling network [38]; and iii) 10 manually curated signaling pathways for cancer from the Cancer Cell Map (http://cancer.cellmap.org/cellmap/). To construct a corresponding human interaction network for both gene expression datasets, we extracted available interactions among common genes in four datasets. Totally, we found 63,113 protein-protein interactions, 1,789 protein-DNA interactions and 3,862 signaling interactions among 10,650 common genes in four datasets.

2.2 Module biomarker identification

To detect the modules, we first extracted the significant network motifs in the integrated cellular network as previously described. Network motifs 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 [39, 40]. They are the smallest basic functional and evolutionarily conserved units in the biological network. Cancer-related genes have also been shown to be more conserved compared to other genes along evolution [41, 42]. We assume that network motifs in a biological network are enriched in “disease driver” genes which are more conserved than other downstream “passenger” genes. These network motifs could form large aggregated modules that perform specific functions by forming collaborations among a large number of network motifs. In this work, we focused on three-node network motifs since larger size network motifs (number of nodes > 3) are composed of three-node ones in most cases [43]. All connected subnetworks containing three nodes in the interaction network were collated into isomorphic patterns, and the number of times each pattern occurred was counted. If the number of occurrences was at least five and statistically significantly higher than random networks, then the pattern was considered as a network motif. The significance test was performed by generating 1000 random networks and computing the fraction of random networks in which the pattern appeared at least as often as in the interaction network. A pattern with $P \leq 0.05$ was considered statistically significant.

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

$$G_k = \left\{ g_{ij} \middle| i=1,2,\ldots,M,\ j=1,2,\ldots,N \right\} \quad (2)$$

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 μ = 0 and standard deviation **σ** = 1 over all samples *j*. The *z* score is defined by

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

where $\widehat{\mu}_i$ is mean expression value of gene *i* across samples, and $\widehat{\sigma}_i$ 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*:

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

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 [44].

The activity score of a subnetwork *A* is then calculated by combining the transformed *z* scores derived from the expression of its individual genes. The individual z_{ij} of each member gene in one subnetwork are combined into the activity of a Z_{A_j} by

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

where w_i denotes the weight that is defined as

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

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):

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

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 network motifs. 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 [45]. 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 network motif 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 network motif was calculated over 100000 random trials in which the expression vectors of individual genes were permuted over the network. The score of each network motif was indexed on the "global" null distribution of all random network motif activity scores. In this study, significant network motifs were selected that have both permutation test *P* values less than 0.0001.

The network motifs that passed the significance tests were clustered in the network motif 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 network motif 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 network motif 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 network motifs.

2.3 Ensemble classification evaluation

After the module biomarkers are identified their reliability is evaluated across different datasets. An ensemble strategy is proposed to increase the stability of our feature selection algorithm, which is a wrapper approach that combines colony optimization with support vector machine (ACO-SVM). The following section describes this feature selection method and evaluation of the method on the basis of classification performance.

2.3.1 Ant colony optimization—Ant colony optimization (ACO) studies artificial systems that takes inspiration from the behavior of real ant colonies [46]. The basic idea of ACO is that a large number of simple artificial agents are able to build good solutions to solve hard combinatorial optimization problems via low-level based communications. Real ants cooperate in their search for food by depositing chemical traces (pheromones) on the ground. Artificial ants cooperate by using a common memory that corresponds to the pheromone deposited by real ants. The artificial pheromone is accumulated at runtime through a learning mechanism. Artificial ants are implemented as parallel processes whose role is to build problem solutions using a constructive procedure driven by a combination of artificial pheromone and a heuristic function to evaluate successively constructive steps.

In this paper, we propose to use ACO for feature selection because of its efficiency and capability in identifying a set of interacting variables that are useful for classification. Also, ACO allows the integration of prior information into the algorithm for improved feature selection.

Through the probability function given below, each ant picks n sets of distinct features from L candidate features:

$$P_i(t)=\frac{(\tau_i(t))^{\alpha}\eta_i^{\beta}}{\sum_i(\tau_i(t))^{\alpha}\eta_i^{\beta}} \quad (8)$$

where $\tau_i(t)$ is the amount of pheromone trail at time t for the feature represented by index i; η_i, represents prior information (e.g. univariate t-statistic) for the feature represented by index i; α and β are parameters that determine the relative influence of pheromone trail and prior information.

At $t = 0$, $\tau_i(t)$ is set to a constant for all features. Thus, at the first iteration, each ant chooses n distinct features (a trail) from L features with probabilities proportional to the existing prior knowledge. Let S_j be the jth ant consisting of n distinct features. Depending on the performance of S_j, the amount of pheromone trail for S_j is updated. The performance function is evaluated on the basis of disease state classification capability of each S_j. We use the features in S_j to build a classifier and estimate the classification accuracy through the cross validation (CV) method. The amount of pheromone trail for each feature in S_j is updated in proportion to the corresponding classification accuracy using

$$\tau_i(t+1)=\rho*\tau_i(t)+\Delta\tau_i(t) \quad (9)$$

where ρ is a constant between 0 and 1, representing the evaporation of pheromone trails; $\Delta\tau_i(t)$ is an amount proportional to the classification accuracy of S_j. $\Delta\tau_i(t)$ is set to zero if the ith feature $f_i \notin S_j$. This update is made for all N ants (S_1,S_2,…,S_N). Note that at t=0, $\Delta\tau_i(t)$ is set to zero for all features. The updating rule allows trails that yield good classification accuracy to have their amount of pheromone trail increased, while others gradually evaporate. As the algorithm progresses, features with larger amounts of pheromone trails and strong prior information influence the probability function to lead the ants towards them.

Compared to particle swarm optimization (PSO) that we previously used in [47] and which is mostly used for continuous optimization problem, ACO is more suitable for discrete optimization problem due to the following reasons: (1) ACO is driven by two parameters: heuristic value and pheromone value. Mostly these values are derived from parameters having discrete values. (2) PSO is driven by neighbor's velocity, which is a continuous parameter as one of the parameters used for deriving velocity is time. Thus, ACO fits better at graph searching problems while PSO fits better at parameter optimization in patter recognition algorithms, because parameters used for graph searching are mostly discrete parameters whereas parameters used for learning/recognition are continuous parameters.

2.3.2 Support vector machines—Support vector machines (SVMs) are learning kernel-based systems that use a hypothesis space of linear functions in high-dimensional feature spaces [48]. In classification problems that involve two classes, linear SVMs search for the optimal hyperplane that maximizes the margin of separation between the hyperplane and the closest data points on both sides of the hyperplane. Thus, parameters of SVMs are determined on the basis of structural risk minimization, not error-risk minimization. Thus, they have the tendency to overcome the overfitting problem. In high dimensional data classification problems, SVMs have proven themselves as one of the pattern classification algorithms with great generalization ability. We will use a linear SVM as the reference classifier for feature selection in module space.

2.3.3 ACO-SVM feature selection algorithm—Ant colony optimization-support vector machine combines ACO and SVM to select features that are useful for SVM classification of samples into two groups. ACO starts with a population of N module sets, where each module set consists of a pre-specified number (n) of distinct modules. Each module is selected from a given set of candidate modules (L) based on its probability function described previously in Eq. (8). SVM classifiers are then built for each module set and the performance of the module set in distinguishing the two groups is evaluated through the five-fold cross-validation method. Using Eq. (9), we update the amount of pheromone trail for each module in proportion to the classification accuracy of the module set, in which the module is involved. The goal is to provide those modules that can lead to improved classification accuracy with better probability of being selected in subsequent iterations.

2.3.4 Ensemble feature selection based on ACO-SVM—In order to select robust module biomarkers for classification in 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 [49]. 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 generate a robust module biomarker set in one dataset, we generated slight variations of the original dataset, and aggregated the outputs of the ACO-SVM feature selection method using 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 can be 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 [50].

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 biomarkers for prognosis purposes. In the following, we present our results and comparison to previously proposed methods applied to the same public datasets.

3.1 Biological interpretability of module biomarkers

The collected biological network involved 72,562 three-node network motifs detected using FANMOD tool [51]. Totally, 1017, 752, 696 and 908 network motifs 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 expression analysis, which does not provide biological insight into the identified markers.

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 (Figure 2). 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.

The “disease driver” genes are usually hub genes in the interaction data, i.e., genes with more than ten surrounding genes. We retrieved the existing interactions surrounding hub genes (genes with more than ten interactions) from the collected molecular interaction data. We observed that only 4 out of 23 hub genes showed discriminative potential (P value <0.01) in module biomarkers identified from van de Vijer dataset. However, these hub genes are important biological markers than other members in one module since they are the center to gather its surrounding differential genes into one module biomarkers.

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 Figure 3, the module markers are more reproducible between datasets than individual marker genes selected without network information (e.g. t-test).

3.2 Classification evaluation of module biomarkers

We tested the classification ability of the identified module biomarkers from four datasets using the proposed ACO-SVM approach. 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 shown in Figure 4.

Through the double-validation strategy, we showed that the module biomarkers outperformed the gene biomarkers in all four experiments. This implies that the module biomarkers are more robust across different datasets generated on different platforms.

3.3 Comparison to existing methods

Several studies have been reported to integrate interaction network information and other biological data (e.g., microarray data) for identification of genetic mediators of disease progression [23, 24, 52, 53]. However, only individual interaction layers, such as the transcriptional layer or the protein complex layer, were modeled by these methods. 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 choose 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 Chuang et al. [24] and Lee et al. [23], 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 better classification accuracy in applying the resulting module marker set to classify previously unseen samples.

4. Discussions

In this paper, we introduced a module-based feature selection framework to identify module biomarkers with high reproducibility and classification accuracy. This was accomplished by a novel hybrid feature selection approach that identifies groups of associated genes by incorporating biological network information, called module biomarkers. Different from traditional data-driven group feature selection methods, we identified the “active subnetworks” within biological network context. The motivation is that a disease or clinical response may be viewed as an emergent behavior of biological network that is altered by the complex interplay of genetic and environmental stimuli. Individual genes tend to collaborate to carry out some specific biological function, in which these genes are called a functional module. In our study, we decomposed biological networks into network motifs - statistically over-represented subgraphs. Cancer-associated genes have been shown to be enriched in particular network motif types, called hotspots in the mammalian cellular signaling network [41]. These hotspots are potentially biomarker clusters or drug target clusters for curing cancer. If a cancer-related gene is mutated in one phenotype, this mutation will influence its surrounding interaction partners at a functional module level. Since the cancer-related genes are usually upstream disease “drivers”, they tend not to be highly differentially expressed compared to their downstream “passenger” genes. On the other hand, these disease driver genes are more stable and study-independent than their downstream “passenger” genes across and within patient samples. We hypothesize that these functional modules are expected to have higher stability and reproducibility in unknown samples, since they have disease “drivers” as the hubs surrounded by their downstream “passengers”. Our findings demonstrated that module biomarkers are more enriched with these disease driver genes such as TP53, which could not be identified by gene-based univariate methods (e.g., t statistic). The reliability of the module biomarkers from different dataset were compared to gene-based approach. The overlaps of module biomarkers from different datasets were largely improved, further confirming that the module biomarkers we identified are likely to be involved in cancer related mechanisms.

Also, we introduced in this paper a strategy in which a set of ensemble feature selection methods was applied to improve biomarker stability and classification performance. In module space, the ensemble feature selection methods were combined with a double-validation strategy to select the optimal module biomarkers according to their classification accuracy on independent datasets. The stability of our ensemble feature selection approach was improved compared to non-ensemble method (Figure 4). This is particularly convenient since it corresponds to sizes of practical interest for the design of a diagnosis/predictive model. As high-quality interaction data become available, such hybrid feature selection methods will help us exploit more disease related information in these and other similar datasets available in human diseases.

Acknowledgments

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

Biographies

Yuji Zhang, Ph.D., is Assistant Professor at the Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo College of Medicine. Dr. Zhang's research interests are in the integrative analysis and visualization of high-throughput biological data, statistical modeling of biological networks, and analysis of massively parallel sequencing data.

Jianhua Xuan, Ph.D., is Associate Professor of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University. Dr. Xuan's research interests include computational systems biology, bioinformatics, intelligent computing, information visualization, advanced image analysis, cellular and molecular imaging, and image guided radiation therapy.

Robert Clarke, Ph.D., D.Sc., is Professor of Oncology and Physiology & Biophysics at Georgetown University Medical Center (GUMC), Dean of Research at GUMC, Interim Director of the Biomedical Graduate Research Organization at GUMC, Associate Vice President of GUMC, and Co- Director of the Breast Cancer Program at the Lombardi Comprehensive Cancer Center. Dr. Clarke studies how hormones, growth factors, and other related molecules affect breast cancer, and how breast cancers become resistant to hormonal and cytotoxic chemotherapies. Dr. Clarke has expertise in the fields of estrogens, antiestrogens, aromatase inhibitors, cell signaling, drug resistance, bioinformatics, and signal transduction.

Habtom W. Ressom, Ph.D., is Associate Professor of Oncology at Georgetown University Medical Center and Director of the Genomics and Epigenomics Shared Resource at the Lombardi Comprehensive Cancer Center. Dr. Ressom is interested in cancer biomarker discovery and systems biology research by analysis of omics data.

References

1. Awan A, Bari H, Yan F, Moksong S, Yang S, Chowdhury S, Cui Q, Yu Z, Purisima EO, Wang E. Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. IET Syst Biol. 2007; 1(5):292–297. [PubMed: 17907678]

2. Azuaje F. What does systems biology mean for biomarker discovery? Expert Opinion on Medical Diagnostics. 2010; 4(1):1–10. [PubMed: 23496106]
3. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003; 31(1):248–250. [PubMed: 12519993]
4. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005; 37(4):382–390. [PubMed: 15778709]
5. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol. 1999; 6(3–4):281–297. [PubMed: 10582567]
6. Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H. PDZBase: a protein-protein interaction database for PDZ-domains. Bioinformatics. 2005; 21(6):827–828. [PubMed: 15513994]
7. Boser, BE.; Guyon, IM.; Vapnik, V. Proceedings of the Fifth Annual Workshop on Computational Learning Theory. Pittsburg: 1992. A training algorithm for optimal margin classifiers.
8. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004; 20(3):374–380. [PubMed: 14960464]
9. Brynildsen MP, Collins JJ. Systems biology makes it personal. Mol Cell. 2009; 34(2):137–138. [PubMed: 19394290]
10. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF. A network-based analysis of systemic inflammation in humans. Nature. 2005; 437(7061):1032–1037. [PubMed: 16136080]
11. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. MINT: the Molecular INTeraction database. Nucleic Acids Res. 2007; 35(Database issue):D572–D574. [PubMed: 17135203]
12. Chen J, Yuan B. Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics. 2006; 22(18):2283–2290. [PubMed: 16837529]
13. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007; 3:140. [PubMed: 17940530]
14. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer. 2008; 8(1):37–49. [PubMed: 18097463]
15. Dorigo M, Di Caro G, Gambardella LM. Ant algorithms for discrete optimization. Artif Life. 1999; 5(2):137–172. [PubMed: 10633574]
16. Duan KB, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience. 2005; 4(3):228–234. [PubMed: 16220686]
17. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005; 21(2):171–178. [PubMed: 15308542]
18. Ergun A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ. A network biology approach to prostate cancer. Mol Syst Biol. 2007; 3:82. [PubMed: 17299418]
19. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):531–537. [PubMed: 10521349]
20. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. 2000; 1(2):RESEARCH0003. [PubMed: 11178228]
21. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004; 32(Database issue):D452–D455. [PubMed: 14681455]
22. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002; 18(Suppl 1):S233–S240. [PubMed: 12169552]

23. Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet. 2005; 365(9458):454–455. [PubMed: 15705441]

24. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005; 33(Database issue):D428–D432. [PubMed: 15608231]

25. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007; 25(3):309–316. [PubMed: 17344885]

26. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008; 4(11):e1000217. [PubMed: 18989396]

27. Lim WK, Lyashenko E, Califano A. Master regulators used as breast cancer metastasis classifier. Pac Symp Biocomput. 2009:504–515. [PubMed: 19209726]

28. Loi S, Haibe-Kains B, Desmedt C, Wirapati P, Lallemand F, Tutt AM, Gillet C, Ellis P, Ryder K, Reid JF, Daidone MG, Pierotti MA, Berns EM, Jansen MP, Foekens JA, Delorenzi M, Bontempi G, Piccart MJ, Sotiriou C. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. BMC Genomics. 2008; 9:239. [PubMed: 18498629]

29. Lonning PE, Sorlie T, Borresen-Dale AL. Genomics in breast cancer-therapeutic implications. Nat Clin Pract Oncol. 2005; 2(1):26–33. [PubMed: 16264853]

30. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, Kershenbaum A, Stolovitzky GA, Blitzer RD, Iyengar R. Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science. 2005; 309(5737):1078–1083. [PubMed: 16099987]

31. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003; 31(1):374–378. [PubMed: 12520026]

32. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet. 2005; 365(9458):488–492. [PubMed: 15705458]

33. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002; 298(5594):824–827. [PubMed: 12399590]

34. Narsing S, Jelsovsky Z, Mbah A, Blanck G. Genes that contribute to cancer fusion genes are large and evolutionarily conserved. Cancer Genet Cytogenet. 2009; 191(2):78–84. [PubMed: 19446742]

35. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D. The MIPS mammalian protein-protein interaction database. Bioinformatics. 2005; 21(6):832–834. [PubMed: 15531608]

36. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003; 13(10):2363–2371. [PubMed: 14525934]

37. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 2001; 98(26): 15149–15154. [PubMed: 11742071]

38. Ressom HW, Varghese RS, Drake SK, Hortin GL, Abdel-Hamid M, Loffredo CA, Goldman R. Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics. 2007; 23(5):619–626. [PubMed: 17237065]

39. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J,

Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437(7062):1173–1178. [PubMed: 16189514]

40. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet. 2004; 36(10):1090–1098. [PubMed: 15448693]

41. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. 2002; 31(1):64–68. [PubMed: 11967538]

42. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005; 122(6): 957–968. [PubMed: 16169070]

43. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A. 2005; 102(38): 13544–13549. [PubMed: 16174746]

44. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005; 310(5748):644–648. [PubMed: 16254181]

45. Tourassi GD, Frederick ED, Markey MK, Floyd CE Jr. Application of the mutual information criterion for feature selection in computer-aided diagnosis. Med Phys. 2001; 28(12):2394–2402. [PubMed: 11797941]

46. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002; 347(25):1999–2009. [PubMed: 12490681]

47. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530–536. [PubMed: 11823860]

48. Wang E, Lenferink A, O'Connor-McCourt M. Cancer systems biology: exploring cancer-associated genes on cellular networks. Cell Mol Life Sci. 2007; 64(14):1752–1762. [PubMed: 17415519]

49. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005; 365(9460):671–679. [PubMed: 15721472]

50. Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. Bioinformatics. 2006; 22(9):1152–1153. [PubMed: 16455747]

51. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002; 30(1):303–305. [PubMed: 11752321]

52. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proc Natl Acad Sci U S A. 2004; 101(16):5934–5939. [PubMed: 15079056]

53. Zhang, Y.; Xuan, J.; de Los Reyes, BG.; Clarke, R.; Ressom, HW. Conf Proc IEEE Eng Med Biol Soc (EMBC 2008). Vancouver, British Columbia, Canada: 2008. Network motif-based identification of breast cancer susceptibility genes.

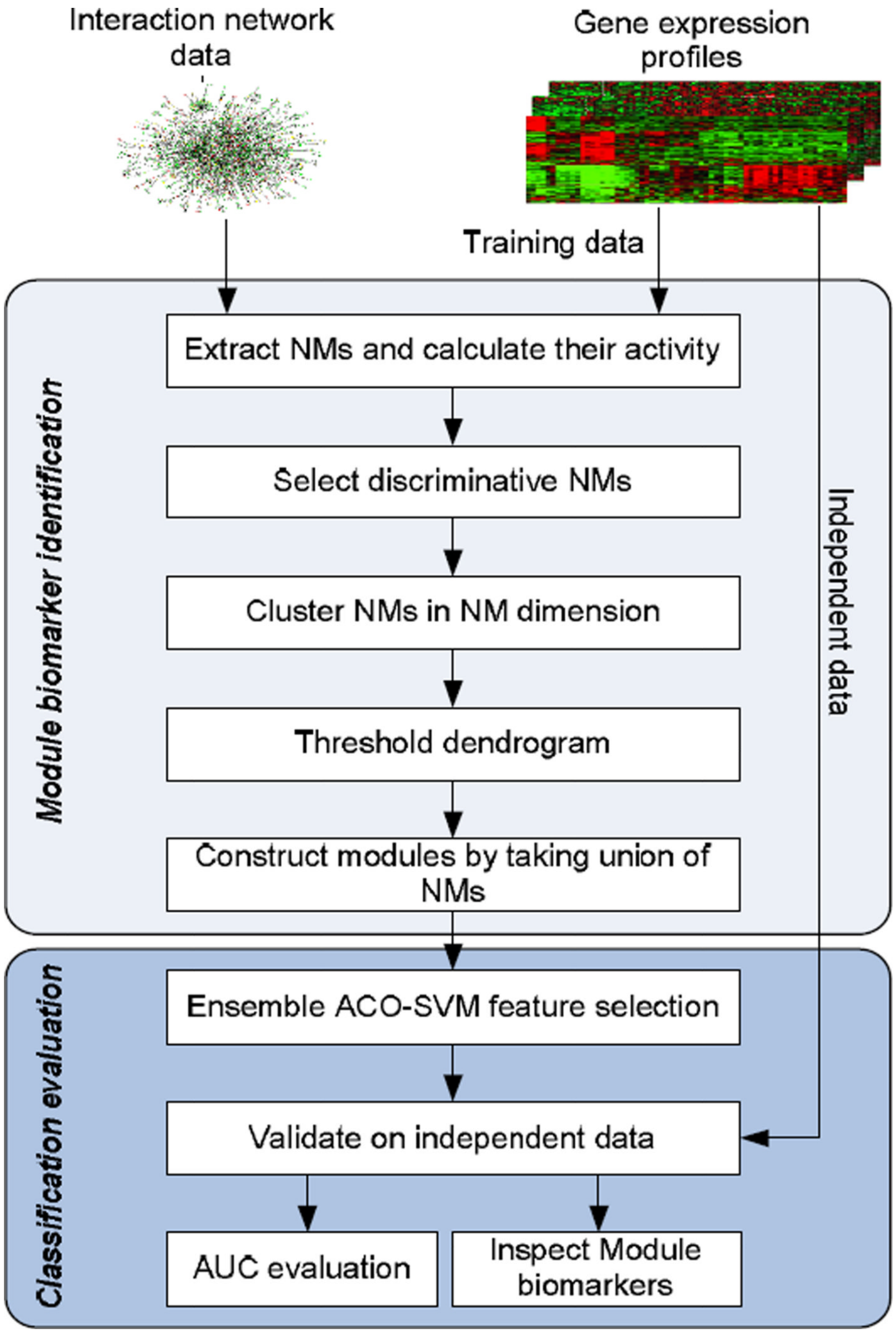
Interaction network data
Gene expression profiles
Training data
Independent data
Module biomarker identification
Extract NMs and calculate their activity
Select discriminative NMs
Cluster NMs in NM dimension
Threshold dendrogram
Construct modules by taking union of NMs
Classification evaluation
Ensemble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

Figure 1.

Schematic overview of module-based biomarker identification and disease classification. NM: network motif; ACO: ant colony optimization; SVM: support vector machine; AUC: area under curve.

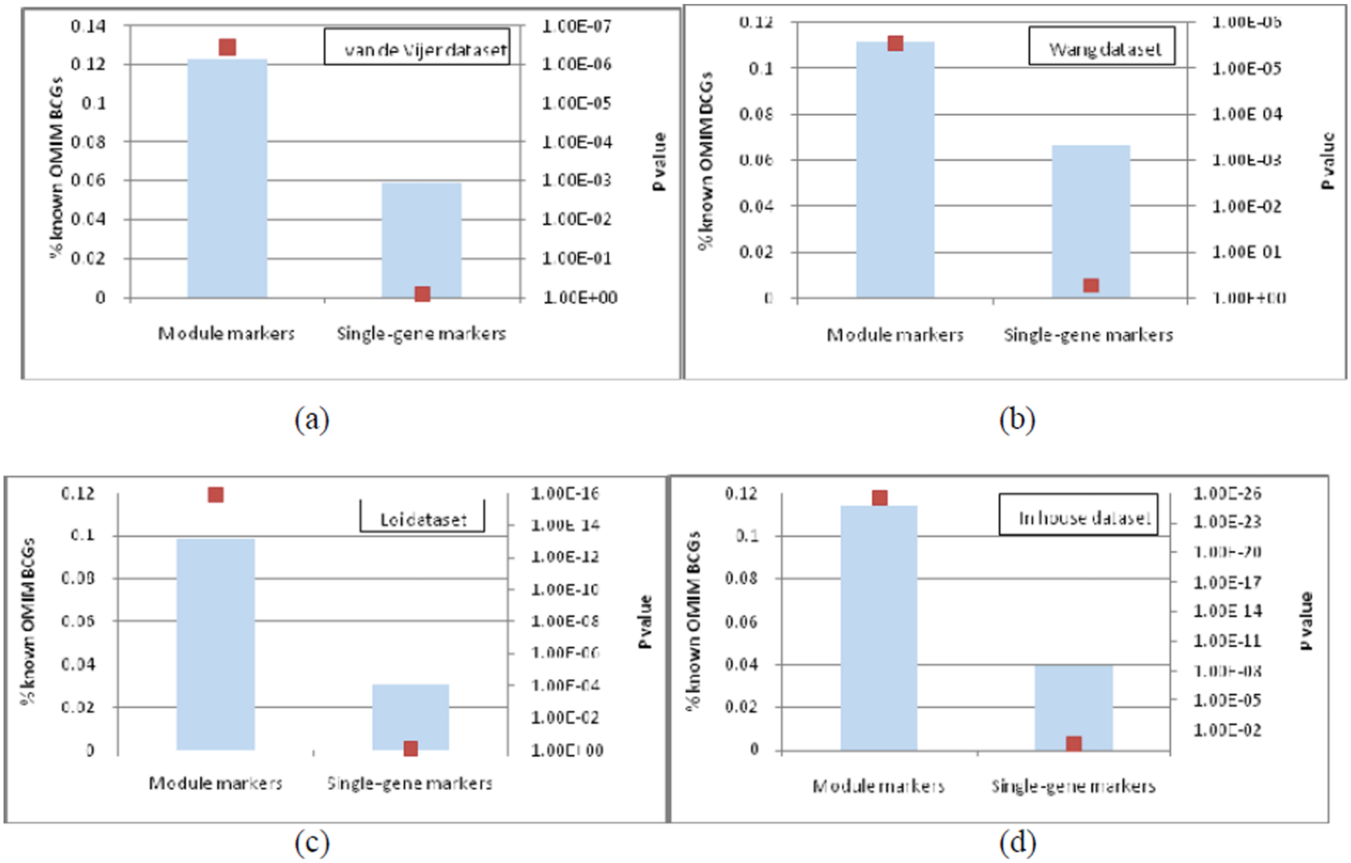

Figure 2.

Detection of BCGs in module biomarkers of four datasets. The enrichment of disease genes is shown for modules or individual genes selected from van de Vijer dataset (a), Wang dataset (b), Loi dataset (c) and our in house dataset (d). Blue bars chart the percentage of BCGs among all genes covered in the markers on the left axis; the red dots chart the hypergeometric *P* values of enrichment on the right axis.

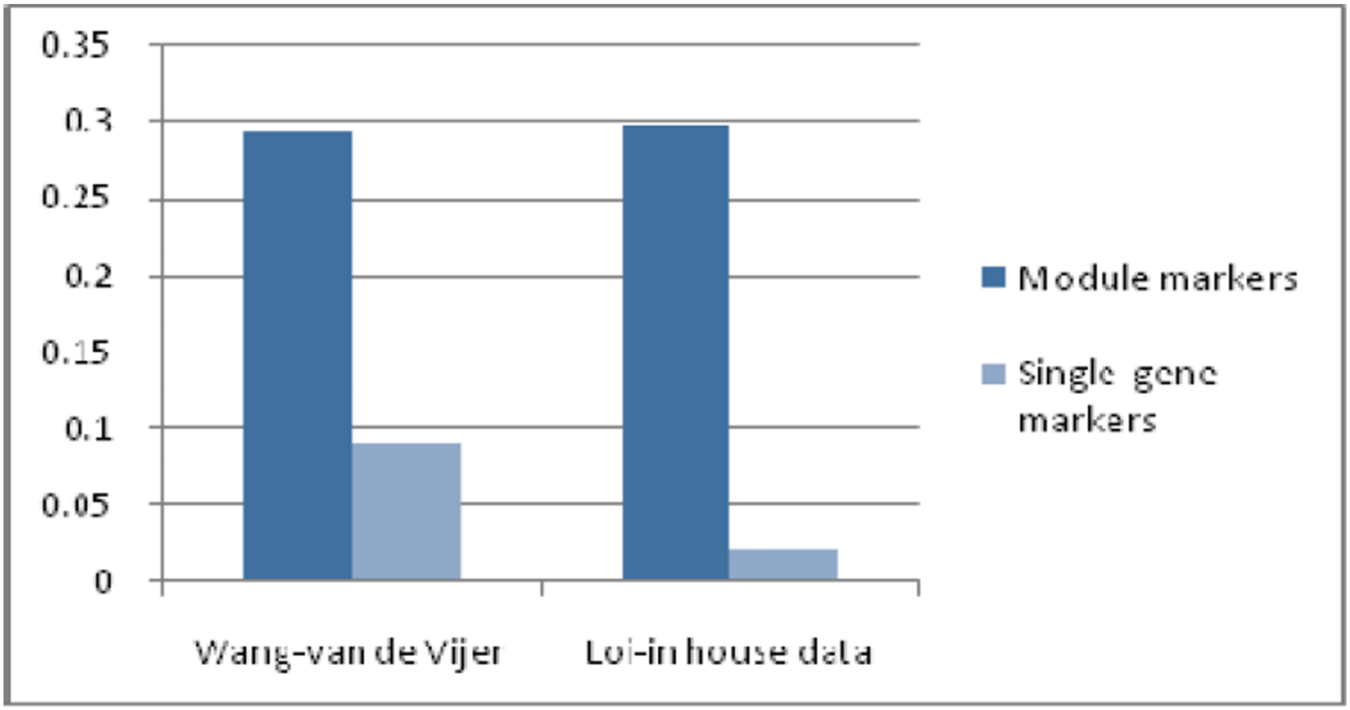

Figure 3.
Agreement in markers selected from one dataset versus those selected from the other dataset in the same clinical group.

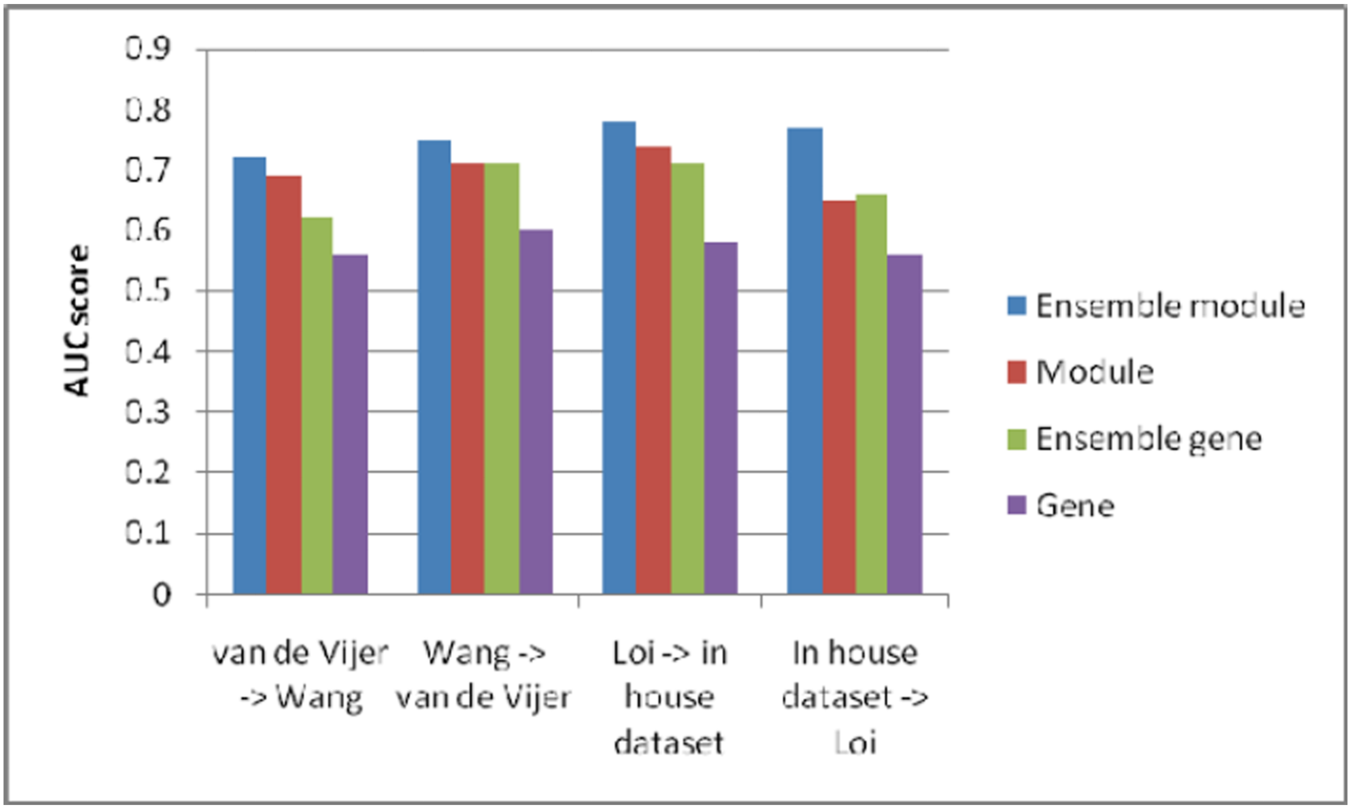

Figure 4.

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

Table 1

Four datasets used for method 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

Table 2

BCGs in module markers derived from four datasets

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