BIOINFORMATICS

Vol. 27 ISMB 2011, pages i205–i213
doi:10.1093/bioinformatics/btr245

Optimally discriminative subnetwork markers predict response to chemotherapy

Phuong Dao[1,†], Kendric Wang[2,3,†], Colin Collins[3,4,*], Martin Ester[1,], Anna Lapuk[3,*,‡] and S. Cenk Sahinalp[1,*,‡]

[1]School of Computing Science, Simon Fraser University, [2]Bioinformatics Training Program, University of British Columbia, [3]Vancouver Prostate Centre and [4]Department of Urology, University of British Columbia

ABSTRACT

Motivation: Molecular profiles of tumour samples have been widely and successfully used for classification problems. A number of algorithms have been proposed to predict classes of tumor samples based on expression profiles with relatively high performance. However, prediction of response to cancer treatment has proved to be more challenging and novel approaches with improved generalizability are still highly needed. Recent studies have clearly demonstrated the advantages of integrating protein–protein interaction (PPI) data with gene expression profiles for the development of subnetwork markers in classification problems.

Results: We describe a novel network-based classification algorithm (OptDis) using color coding technique to identify optimally discriminative subnetwork markers. Focusing on PPI networks, we apply our algorithm to drug response studies: we evaluate our algorithm using published cohorts of breast cancer patients treated with combination chemotherapy. We show that our OptDis method improves over previously published subnetwork methods and provides better and more stable performance compared with other subnetwork and single gene methods. We also show that our subnetwork method produces predictive markers that are more reproducible across independent cohorts and offer valuable insight into biological processes underlying response to therapy.

Availability: The implementation is available at: http://www.cs.sfu.ca/~pdao/personal/OptDis.html

Contact: cenk@cs.sfu.ca; alapuk@prostatecentre.com; ccollins@prostatecentre.com

1 INTRODUCTION

In the treatment of cancers, patients presenting tumors with similar clinical characteristics will often respond differentially to the same chemotherapy (van't Veer and Bernards, 2008). In fact, for many types of cancer, only a minority of treated patients will observe regression of tumor growth. This is the case for both conventional chemotherapeutic agents and newer targeted therapies that affect specific molecules. To achieve an effective cancer treatment, it is critical to identify the underlying mechanisms that confer chemoresistance in some tumors but not others.

The advent of genome-wide expression profiling technologies has allowed the discovery of novel biomarkers for cancer diagnosis, prognosis and treatment (van't Veer and Bernards, 2008). While some progress has been made toward identifying reliable prognostic markers for breast and other cancers, development of molecular markers predictive of response to chemotherapy has proved to be far more difficult (van't Veer and Bernards, 2008).

In recent years, a number of studies have used genome-wide expression profiling to identify genes that could be used as predictors of drug response in breast cancer (Cleator *et al.*, 2006; Hess, 2006). In these studies, single gene marker methods were used, where each gene is individually ranked for differential expression and the top genes were selected as predictors known as single gene markers. Additional study (Lee *et al.*, 2007; Liedtke *et al.*, 2010) required single gene markers not only to be differentially expressed but also to have similar coexpression between the training and test cohorts. While some of these predictive markers have shown promising results in a limited number of patient cohorts, many of these signatures have failed to achieve similar performance in additional validation studies (Bonnefoi *et al.*, 2009). In addition, single gene markers developed from different cohorts have been shown to have very little overlap (Ein-Dor *et al.*, 2006). A further limitation of single gene markers is that they provide relatively limited insight into the biological mechanisms underlying response to drug response. Thus, predictive markers with robust performance, greater reproducibility and improved insights into drug action–which are critical for clinical application–still remains elusive.

Previous studies have observed that gene products associated with cancer tend to be highly clustered in coexpression networks and have more 'interactions'. Inspired by this observation, Chuang *et al.* (2007) introduced the use of all members of a protein–protein interaction (PPI) subnetwork as a metagene marker for predicting metastasis in breast cancer. Chuang *et al.* (2007) demonstrated that subnetwork markers are more robust, i.e. their results tend to provide more reproducible results across different cohorts of patients. Motivated by the limitations in predicting drug response using single gene markers and the better performance promised by subnetwork markers, this article aims to identify subnetwork markers to predict chemotherapeutic response, as detailed below.

1.1 Subnetwork markers in other applications

Chuang *et al.* (2007) defined subnetwork activity as the aggregate expression of genes in a given subnetwork. The discriminative score of a subnetwork—which reflects how well the subnetwork discriminates samples of different phenotypes (or classes)–was derived from mutual information between subnetwork activity and the phenotype. The study presented greedy algorithms for identifying subnetworks with the highest discriminative scores and

*To whom correspondence should be addressed.

†The authors wish it to be known that in their opinion, the first two authors should be regarded as joint First Authors.

‡The authors wish it to be known that in their opinion, the last two authors should be regarded as joint Last Authors.

demonstrates significant improvement in classification performance over single gene marker approaches.

Another approach introduced by Chowdhury and Koyutürk (2010) used a binary representation of gene expression profiles to retrieve subnetwork markers. Binarized gene expression profiles were overlaid on PPI networks and subnetworks that contain genes differentially expressed in all the samples from a given class are chosen as markers. Using this approach, Chowhury *et al*. were able to predict colon cancer metastasis with high confidence. Recently, this group introduced an extension of their previous algorithm which takes into account patterns of differential expression for improved classification performance (Chowdhury, 2010). A similar approach using binary representation of gene expression profiles was published by Ulitsky *et al.* (2008), where subnetworks analysis was applied to the identification of dysregulated pathways in Huntington's disease.

More recently, Su *et al.* (2010) identified paths containing many differentially expressed and coexpressed genes from PPI networks and greedily combined these paths to obtain subnetwork markers for predicting breast cancer metastasis. Wu *et al.* (2010) published a report on the application of a network-based approach to drug response data in Type 2 Diabetes. Samples were expression profiled upon treatment with individual drugs and affected subnetworks for these drugs were retrieved. These subnetworks were used to score the individual drug effect and further used to predict the effectiveness of the combination of two drugs. While the approach by Wu *et al.* (2010) proved to be useful for the prediction of a drug effect on network activities, the study did not attempt to develop markers of response to any given therapy in any individual patient.

Subnetworks from functional association networks can also be used for the development of markers. Edges among gene products in such networks [for example STRING database (Jensen, 2009)] are scored based on integration of different sources of information such as high-throughput experiments, physical binding extracted from literature and coexpression networks built from many microarray experiments. Spirin and Mirny (2003) and King *et al.* (2004) have observed that the constituent gene products of dense subnetworks contain many more edges than expected and usually participate in the same biological function/process or belong to the same protein complex. Dense functional association subnetworks have been used in two recent papers by Dao *et al.* (2010) and Fortney *et al.* (2010), which demonstrate that dense networks significantly improve the performance of subnetwork markers from PPI networks in the classification of colon cancer metastasis and prediction of chronological age in *Caenorhabditis elegans*.

Despite their improved performance, available approaches have a number of disadvantages. Network-based approaches introduced by Chuang *et al.* (2007), Fortney *et al.* (2010) and Chowdhury and Koyutürk (2010) are heuristic methods and thus do not guarantee the optimality of the solution for marker selection—an optimal solution would presumably provide a better prediction performance. The branch and bound approach (Chowdhury, 2010) or exhaustive enumeration (Dao *et al.*, 2010) can yield an optimal solution under some fixed set of parameters; however, their worst-case running time can be super-polynomial (and hence intractable). Therefore, there is a keen need for designing efficient algorithms to retrieve the optimal subnetwork markers that could successfully distinguish samples from different classes.

1.2 Our contributions

In this article, we introduce a novel and efficient randomized algorithm to compute 'optimally discriminative' subnetworks for classification of samples from different classes. The discriminative score is calculated as the difference between the total distance between samples from different classes and the total distance between samples from the same class. Our algorithm is based on the color-coding paradigm (Alon *et al.*, 1995), which allows for identifying the optimally discriminative subnetwork markers for any given error probability. Since the running time of our algorithm is a logarithmic function of the error probability, we can set the error probability to a small value, close to zero, while the running time does not increase much. When the maximum size of a subnetwork is $k = O(\log n)$, where n is the size of the network, we have a polynomial time algorithm with a fixed error probability.

The evaluation of our method on published patients' drug response data demonstrate that optimally discriminative subnetwork markers yield both greater and more robust classification performance compared with single gene markers and other subnetwork markers. Moreover, our algorithm provides classification results which are reproducible across independent cohorts, and provide greater biological interpretation of the underlying mechanisms of chemotherapy response.

Since the discriminative score is additive, we can easily adapt our method to retrieve subnetwork markers to distinguish samples from more than two classes. This is very helpful, in particular, when there are more than three categories for responses to treatment: complete, partial and non-response.

2 METHODS

In our methodology, each patient sample is represented as a point in high-dimensional space where each dimension represents one gene. We perform dimensionality reduction by projecting samples (points) into a subspace of at most k dimensions such that samples from different classes are well separated. The separation criteria is defined based on minimizing the distances of samples from the same class while maximizing the distances of samples from different classes. Figure 1 sketches the idea behind our approach.

We formalize our problem as the Optimal Discriminating k-Subnetwork (ODkS) problem below. We then assess the complexity of the problem and finally give a randomized algorithm to solve it for any given error probability.

2.1 Problem definition and its complexity

Before formally defining ODkS problem, we would like to introduce the notations used. Without loss of generality, we assume that we have only two classes of samples: positive and negative. Note that it is easy to generalize our approach for more than two classes. Let A and A' denote the expression matrices for positive and negative samples, respectively. For each gene g_i, let A_i and A'_i, respectively, denote the expression profiles of gene g_i in positive class and negative class. For expression matrix A (A'), let $A_i(j)$ ($A'_i(j)$) denote the expression of g_i in sample j.

Given n genes, let a and a' denote the number of samples in positive class and negative class, respectively. We denote the PPI network by $G=(V,E)$, where $|V|=n$ and $|E|=m$.

We define the weight function w on subnetwork S as the difference between the total distance between samples from different classes and the

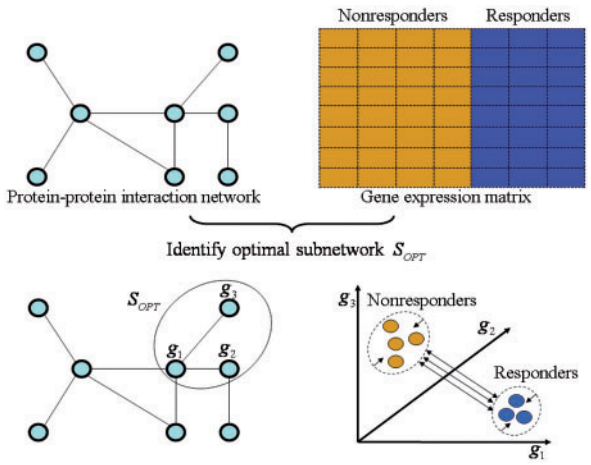

Fig. 1. The main idea behind our approach: samples (denoted as points in a high-dimensional space) are projected into k-dimensional space while ensuring that samples from the same class are clustered together, while samples from different classes stay separated. These k dimensions/genes have to form a connected subnetwork in a PPI network. The main difference between our approach and earlier ones is that we can identify the optimal subnetwork S_{OPT} in polynomial time when $k=O(\log n)$; here n is the size of the network. This is done by minimizing the total distance of samples from same class while maximizing the total distance of samples from different classes.

total distance between samples from the same class—under L_1 distance:

$$w(S) = \sum_{j=1}^{a}\sum_{j'=1}^{a'}\sum_{\forall i:g_i\in S}\frac{|A_i(j)-A_i'(j')|}{aa'} - \sum_{j=1}^{a}\sum_{j'=1}^{a}\sum_{\forall i:g_i\in S}\frac{|A_i(j)-A_i(j')|}{aa} - \sum_{j=1}^{a'}\sum_{j'=1}^{a'}\sum_{\forall i:g_i\in S}\frac{|A_i'(j)-A_i'(j')|}{a'a'}$$

The ODkS problem asks to compute the connected subnetwork S_{OPT} ($|S_{\mathrm{OPT}}|\le k$) from G such that S_{OPT} 'distinguishes' samples from different classes 'optimally', i.e. $w(S_{\mathrm{OPT}})$ is the maximum among $w(S)$'s for any connected subnetwork S. We call S_{OPT} the optimally discriminative subnetwork.

For any connected subnetwork S, $w(S)$ could be rewritten as:

$$w(S)=\sum_{\forall i:g_i\in S}\left(\sum_{j=1}^{a}\sum_{j'=1}^{a'}\frac{|A_i(j)-A_i'(j')|}{aa'} - \sum_{j=1}^{a}\sum_{j'=1}^{a}\frac{|A_i(j)-A_i(j')|}{aa} - \sum_{j=1}^{a'}\sum_{j'=1}^{a'}\frac{|A_i'(j)-A_i'(j')|}{a'a'}\right)$$

We will extend the discriminative score function w so that it can apply on a single gene. We assign each gene g_i to a weight $w(g_i)$:

$$w(g_i)=\sum_{j=1}^{a}\sum_{j'=1}^{a'}\frac{|A_i(j)-A_i'(j')|}{aa'} - \sum_{j=1}^{a}\sum_{j'=1}^{a}\frac{|A_i(j)-A_i(j')|}{aa} - \sum_{j=1}^{a'}\sum_{j'=1}^{a'}\frac{|A_i'(j)-A_i'(j')|}{a'a'}$$

Now we can rewrite the discriminative score of a connected subnetwork S as:

$$w(S)=\sum_{\forall i:g_i\in S} w(g_i)$$

Thus, identifying the optimally discriminative connected subnetwork S_{OPT} ($|S|\le k$) is equivalent to finding the connected subnetwork for which the total weight of the vertices is maximum possible. A variant of this problem without any restriction on the size of the subnetworks ($k\le n$) was defined to extract dysregulated pathways in different cancer types by two independent studies (Dittrich *et al.*, 2008; Qiu *et al.*, 2009). Both studies provided integer linear programming formulations but rather than solving the IP formulation, Qiu *et al.* (2009) solved a relaxed version of the program, thus, did not give the optimal solution, and Dittrich *et al.* (2008) tried to solve the integer linear program using a cutting plane method—however, this approach does not guarantee a worst-case running time.

Another variant of the ODkS problem, the Connected k-Subgraph problem (where the weights of vertices are either 0 or 1), is proved to be NP-hard by (Hochbaum,D.S. and Pathria, A., in press). Here we prove that ODkS problem is also NP-hard:

THEOREM 2.1. *The ODkS problem is NP-hard even when we have one sample for each class.*

PROOF. The reduction is done from Connected k-Subgraph problem defined by (Hochbaum,D.S. and Pathria, A., in press). We are given an instance of Connected k-Subgraph problem where we have a graph $G=(V,E)$, a weight function $h:V\longrightarrow\{0,1\}$ and positive integers k and l. For a subnetwork S, let $g(S)$ be the number of vertices with weight 1. The Connected k-Subgraph problem asks whether there exists a subgraph S in G with at most k vertices such that $g(S)\ge l$.

We build an instance of the ODkS problem as follows. The network G' for the instance of ODkS problem is the same as the given graph G i.e. $V'=V$ and $E'=E$. We only have one sample $a=1$ for the positive class and another sample for the negative class $a'=1$. For each gene g_i corresponding to a vertex v_i in G, set $A_i(1)=0$ and $A_i'(1)=h(v)$. Now for every vertex v_i such that $h(v_i)=1$, we have the discriminative score $w(v_i)=1$. By the construction, the discriminative score of any connected subnetwork S' from G' is equivalent to the number of vertices with weight 1 of the corresponding subgraph S in G. Thus, G has a subgraph S with at most k vertices and $g(S)\ge l$ if and only if the network G' has a subnetwork S' also with at most k vertices and $w(S)\ge l$.

2.2 A randomized algorithm

In this section, we give a randomized algorithm to solve the ODkS problem for any given error probability. This randomized algorithm is based on color-coding technique (Alon *et al.*, 1995).

Color coding is an algorithmic technique that was first introduced by Alon *et al.* (1995) to detect a simple path or a cycle of length k in a given graph. The algorithm consists of a predefined number of iterations. In each iteration, there are two main steps: assign each vertex uniformly at random with one of k colors and detect whether there is a 'colorful' path or cycle of length k in the given graph. A path or cycle is colorful if it is not the case that two vertices u,v in the path or cycle have the same color.

The idea behind the algorithm is the clever use of colors to reduce the number of paths that need to consider in the detecting step. In the naive algorithm, we need to keep track of every vertices visited so far which uses $O(n^k)$ time and space. Now we only keep track of all possible sets of vertices of distinct colors which only take $O(n2^k)$ time and space.

Color coding is widely applicable in the context of retrieving 'homologous' subnetworks from a PPI network given a particular query pathway or protein complex (Bruckner *et al.*, 2010; Dost *et al.*, 2008; Scott *et al.*, 2006; Shlomi *et al.*, 2006). Color coding has also been successfully applied to retrieve network motifs (subnetworks which are recurrent more than expected in a PPI network) and comparing PPI networks of different species (Alon *et al.*, 2008; Dao *et al.*, 2009).

Similar to color-coding technique, our algorithm consists of a predefine number of iterations n_i (we will show how to determine n_i later). Each iteration consists of two main steps:

(1) Assign a vertex uniformly at random with one of k colors.

(2) Identifying the colorful connected subnetwork S'_{OPT} ($|S'_{\mathrm{OPT}}| \leq k$) with the maximum discriminative score $w(S'_{\mathrm{OPT}})$.

We remind the readers that S_{OPT} is the optimally discriminative connected subnetwork while S'_{OPT} is the colorful optimally discriminative subnetwork after each iteration. After n_i iterations, we return S'_{OPT} of some iteration that has the the maximum $w(S'_{\mathrm{OPT}})$. We will prove that we return S_{OPT} with the given error probability δ by determining the number of iterations n_i and identifying the colorful optimally discriminative subnetwork S'_{OPT} in the second step efficiently.

In the following, we describe how to estimate the number of iterations n_i. For each iteration, the probability that we could retrieve S_{OPT} is the same as the probability that S_{OPT} is colorful which is $k!/k^k \geq e^{-k}$. In order to boost the success probability to at least $1-\delta$ for a given error probability δ, we need

$$n_i \leq ln(1/\delta)e^k$$

iterations to yield the S_{OPT}.

In what follows, we describe an efficient dynamic programming algorithm to retrieve the S'_{OPT}. At each iteration, for any vertex $v \in V$ let color(v) denote the color of v. By extending the notation of the discriminating function w defined earlier, we let $w(u,T)$ denote the colorful connected subnetwork S' such that S' contains u, the color set of vertices in S' is T and S' has the maximum discriminative score compared with ones of the colorful connected subnetwork S'''s that contain u. For the base case, for each vertex u, we have:

$$w(u,\{c\}) = \begin{cases} w(u) & \text{if } c = \text{color}(v) \\ -\infty & \text{otherwise.} \end{cases}$$

In the general case, we can compute $w(u,T)$ as follows:

$$w(u,T) = \max_{\forall v: uv \in E} \left\{ \max_{\forall P,Q: P \cap Q = \emptyset, P \cup Q = T} \{w(u,P) + w(v,Q)\} \right\}$$

Here we assume that the addition of $-\infty$ and any real number or $-\infty$ is $-\infty$. We first compute $w(v,T_1)$ for each vertex v and each set T_1 of one color and so on. In the final step, we compute $w(v,T_k)$ for each vertex v and each set T_k of k colors. Now we compute $w(S'_{\mathrm{OPT}})$ as follows:

$$w(S'_{\mathrm{OPT}}) = \max_{\forall v: v \in V} \left\{ \max_{\forall T: T \neq \emptyset, |T| \leq k} \{w(v,T)\} \right\}$$

Now we estimate the running time complexity of this randomized algorithm. Let deg(u) be the degree of vertex u. For any vertex u and a set of colors T, in order to compute each $w(u,T)$, it takes $O(\deg(u)2^{|T|})$ time. To retrieve S'_{OPT} at each iteration, it takes $O(mk4^k)$ time. Thus, the worst-case running time to retrieve S_{OPT} is $O(mk \ln 1/\delta (4e)^k)$. For our interests in subgraphs of small size $k = O(\log n)$ and for a fixed probability of error, it takes polynomial time to find the optimally discriminative subnetwork S_{OPT}.

2.3 Ranking subnetwork markers

From now on, we fix the error probability $\delta = 0.001$ and the maximum size of a subnetwork $k = 7$ of for any experiment performed later. For each vertex $v \in V$ and for each size n' from 4 to k, we compute the optimal discriminative subnetwork that contain v with n' vertices. In total, we have at most kn subnetworks.

For each subnetwork S, we aggregate the expression profiles of genes in S into a metagene s:

$$A_s(j) = \sum_{g_i \in S} A_i(j)/|S| \ (1 \leq j \leq a)$$

$$A'_s(j') = \sum_{g_i \in S} A'_i(j')/|S| \ (1 \leq j' \leq a')$$

Now the normalized discriminative score of a subnetwork S is calculated in the same way as we calculate the discriminative score $w(g)$ for any gene g in Section 2.1. We rank all the extracted subnetworks by their normalized discriminative score. Then we select subnetwork markers from the top to the bottom of the list as follows. Suppose L is the number of genes in the selected subnetworks so far and S is the current considered subnetwork. S is only selected if we have at least $|S|/2$ genes that are not from L. We finish the selection process with 50 subnetworks.

2.4 Classification process and performance assessment

We always consider top 50 subnetworks for our method for any experiment performed after this point. For any l ($1 \leq l \leq 50$), we represent a sample using top l subnetworks ($S_1, \ldots, S_l$) as follows. Each sample j is transformed into a l-dimensional vector $V(j) \in \mathbb{R}^l$ where the entries $V(j)_l$ for each marker l are

$$V(j)_l := \sum_{v \in S_l} E(v,j)/|S_l|$$

where v ranges over all genes v contained in the subnetwork marker S_l and $E(v,j)$ is the expression of gene v in sample j. In other words, each sample j becomes a point $V(j)$ in the l-dimensional feature space $\mathbb{R}^l$. Now, all the classification experiments were performed using three-nearest neighbor classifier under L_1 distance.

Since the tested datasets have an imbalanced ratio between number of samples in positive and negative class, accuracy is not a good measure for classification performance. We utilize Matthews Coefficient Correlation (MCC) as a measure to compare different classifiers (BW, 1975). MCC is essentially the Pearson correlation between the vectors of predicted labels and true labels of a testing set. Suppose that TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. The MCC can be also calculated as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

If one of the four sums in the denominator is zero, the denominator is set to one. This results in a Matthews correlation coefficient of zero. MCC value of 1 indicates a perfect prediction, -1 an inverse prediction and 0 a completely random prediction.

MCC is a recommended measure when compared with other measures for classification performance (Baldi, 2000). We have chosen MCC over area under ROC curve (AUC) to facilitate comparison to competing models from the MAQC-II study (Shi, 2010).

3 RESULTS AND DISCUSSION

3.1 Dataset and network

We retrieved the human PPI data from the Human Protein Reference Database (HPRD) version April 2010 (Prasad *et al.*, 2009). By including both binary interactions and considering each protein complex as a clique of proteins, we obtained 46 370 protein interactions involving 9617 proteins.

We assessed the performance of our method on a human breast cancer dataset contributed by the University of Texas M.D. Anderson Cancer Center (MDACC, Houston, TX, USA). The gene expression profiles were retrieved from NCBI Gene Expression Omnibus (GEO) with accession number GSE20194. Gene expression data from 230 Stage I–III breast cancers were generated from fine-needle aspiration specimens of newly diagnosed breast cancers before any therapy. Patients received 6 months of neoadjuvant chemotherapy comprising paclitaxel (T), 5-fluorouracil(F), doxorubicin (A) and cyclophosphamide (C) (and denoted as TFAC) followed by surgical resection of the cancer. Responders to chemotherapy was categorized as a pathological complete response i.e. no residual invasive cancer in the breast or lymph nodes or residual

invasive cancer. RNA extraction and gene expression profiling were performed in multiple batches over time using Affymetrix U133A microarrays. This dataset was split into two different cohorts according to the time of collection. One cohort consists of 130 samples while the other one consists of 100 samples. The expression profiles were normalized with Robust-chip Median Average (RMA) algorithm (Irizarry *et al.*, 2003) and adjusted for batch effect using ComBat (Johnson *et al.*, 2007). Prior to model generation, the expression values of the two cohorts were normalized but not standardized.

3.2 Classification performance

We evaluated the performance of our method (we denote as OptDis) against both single gene marker models and other subnetwork-based methods following the workflow presented by the MicroArray Quality Control (MAQC)-II studies (Popovici, 2010; Shi, 2010). In those studies, the MAQC project assessed the performance and limitations of various data analysis methods in developing and validating microarray-based predictive models with the ultimate goal of discovering best practices. Thirty-six groups participated in the project to develop classifiers for 13 large datasets, including the one used in our study. MAQC models (denoted as MAQC) were constructed by these groups using different methods for data processing (i.e. normalization), feature selection and classification.

To assess the predictive performance, we performed two analyses. In the forward cross-dataset (FXD) analysis, we treated the 130 patient cohort as the training set used for deriving markers, and validated their performance on the 100 patient cohort. We also performed the complementary backward cross-dataset (BXD) analysis and swapped the cohorts used in training and validation. In Figure 2, we compare the performance of OptDis against single gene marker models. The single gene marker classifier constructed using *t*-test is denoted by SGM and includes only genes that map to the PPI network. For each mappable gene, the corresponding probe with the lowest *P* value was used in the model. We also compared the performance of our method OptDis against implementations of existing subnetwork-based methods, one based on mutual information (GreedyMI) (Chuang *et al.*, 2007), and another based on dense subnetworks (we denote as Dense) using the STRING functional network (Dao *et al.*, 2010). The density threshold to extract all dense subnetworks is set at 0.7 as implemented in Dao *et al.* (2010). Note that, top 50 subnetworks for GreedyMI and Dense are ranked based on their mutual information scores. Starting from around 20 features, the performance of OptDis is better than competing methods. While the maximum MCC value is not that high, it is still significant compared with the random classifier which has an MCC value of 0. Moreover, predicting response to chemotherapy has been shown as a difficult endpoint to predict in the recent MAQC publications (Shi, 2010). The difficulties might be due to the known heterogeneity within tumors of the same cancer type, subtype-specific response, differences in drug metabolism between individuals and variations in chemotherapy schedules between patients (Popovici, 2010). Figure 3 shows the average performance of models in cross-dataset validation of FXD and BXD analyses. Here, the average performance for a model is the average MCC of 50 models generated using the top 1–50 features. The MAQC performance was derived from the average of top model from each participating group. As shown in Figure 3, OptDis

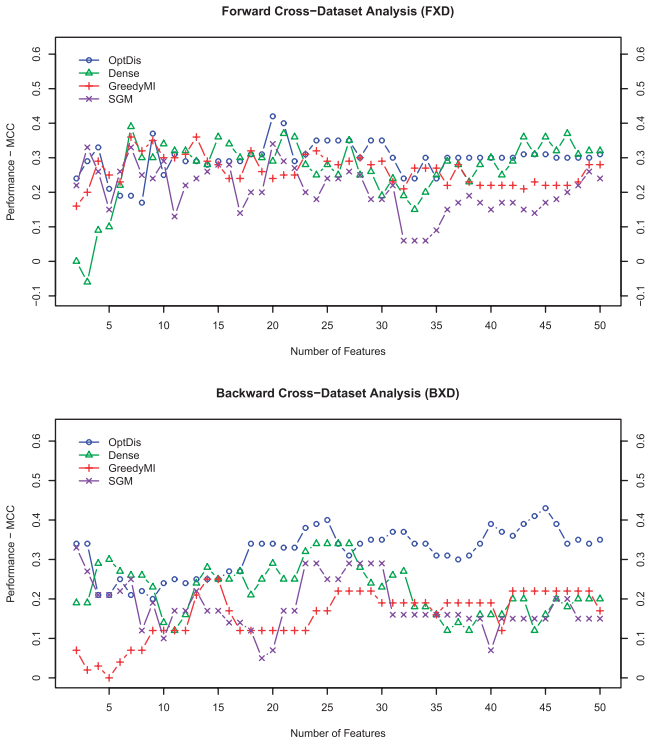

Fig. 2. Line graphs show the MCCs for different predictive models using the top 1–50 features. The compared approaches are single gene marker model based on *t*-test (SGM) and subnetwork marker models include Chuang *et al.* (2007) (GreedyMI), dense subgraphs from STRING functional network by Dao *et al.* (2010) (Dense) and our approach (OptDis).

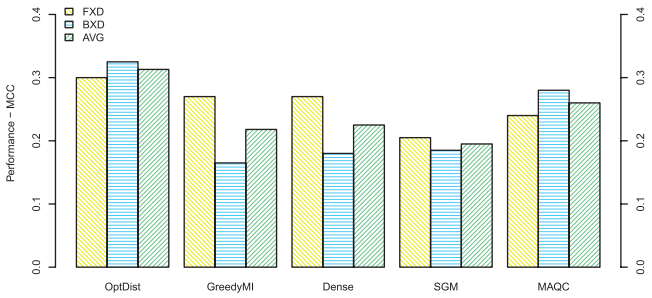

Fig. 3. Bar charts show the average MCCs of different predictive models. Single gene marker models include one based on *t*-test (SGM) and models from MAQC project (MAQC). Subnetwork marker models include Chuang *et al.* (2007) (GreedyMI), Dao *et al.* (2010) (Dense), and our method (OptDis). The yellow bars and blue bars show the classification performance in FXD and BXD analyses respectively. The green bars show the overall average performance, calculated as the average of the yellow and blue bars.

outperforms all the other competitors on the average classification performance in FXD and BXD analyses.

For further analyses, we compared the average best performance of different classifiers in Figure 4. The average best performance of a classifier is the average of its best model from FXD analysis and its best one from BXD analysis. Here, we compare against the top three MAQC models. Figure 4 shows that our top OptDis model

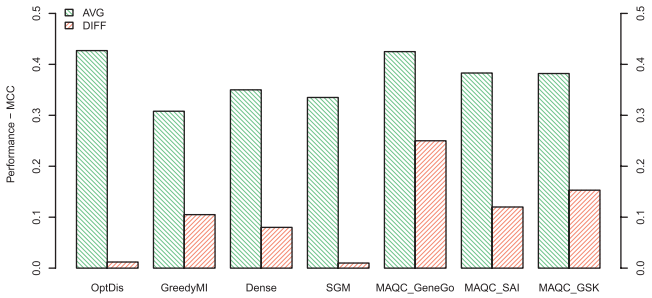

Fig. 4. The bar charts show the average best MCCs of different classifiers. The average best performance of a classifier is the average of its best model from FXD analysis and its best model from BXD analysis. Single gene marker models include one based on *t*-test (SGM) and top 3 models from MAQC project (MAQC GeneGo, MAQC SAI, MAQC GSK). Subnetwork marker models include Chuang *et al*. (2007) (GreedyMI), Dao *et al*. (2010) (Dense), and our approach (OptDis). Green bars show the average best MCCs and red bars show the difference in MCC between a classifiers' best model from FXD analysis and its best model from BXD analysis.

has consistent performance in cross-dataset validation experiments. In contrast, the top three MAQC models show discrepancy in performance when the datasets used for training and test were swapped—especially in the case of the MAQC_ GeneGo model, which has the largest difference in performance (0.25) between the FXD and BXD analysis. The second and third best MAQC models also show similar discrepancy in performance.

Figure 5 shows the performance of OptDis against one of the predictive model constructed, where the constituent genes were taken from the top x ($1 \leq x \leq 50$) OptDis subnetworks (we denote as SGM_OptDis). We also compare our method against another single gene marker model that ranks all genes by *t*-test and matches the number of genes in the top x ($1 \leq x \leq 50$) subnetworks from OptDis (SGM_M). OptDis is consistently better than SGM_OptDis across different number of features. This suggests the importance of treating genes as functional modules. Moreover, on the average constituent genes taken from OptDis subnetworks tend to perform better than genes from simple predictive model using *t*-test. Hence, OptDis subnetworks might capture genes more informative to predicting chemotherapy response.

In summary, our subnetwork markers have the best combination of relatively high performance and greater stability between different cohorts of patients and thus could be more clinically applicable to other independent cohorts of patients.

3.3 Reproducibility of predictive markers

We compared the reproducibility of predictive markers derived from subnetwork (SN) and single gene (SG) approaches by training OptDis and SGM on the two different cohorts of 130 and 100 breast patients and calculating the number of overlapping genes. For this comparison, we considered the top 50 SN markers (T50 SN), the top 50 SG markers (T50 SG) and the top X SG markers (Tx SG) that comprises a similar number genes as the T50 SNs. There is an overlap of 39 genes (27–30%) between the T50 SN, significantly more than the overlap of five genes (10%) between the T50 SG and 25 (17%) genes between T150 SG. Greater reproducibility may be contributing to the improved stability in predictive performance of our subnetwork markers over single gene markers.

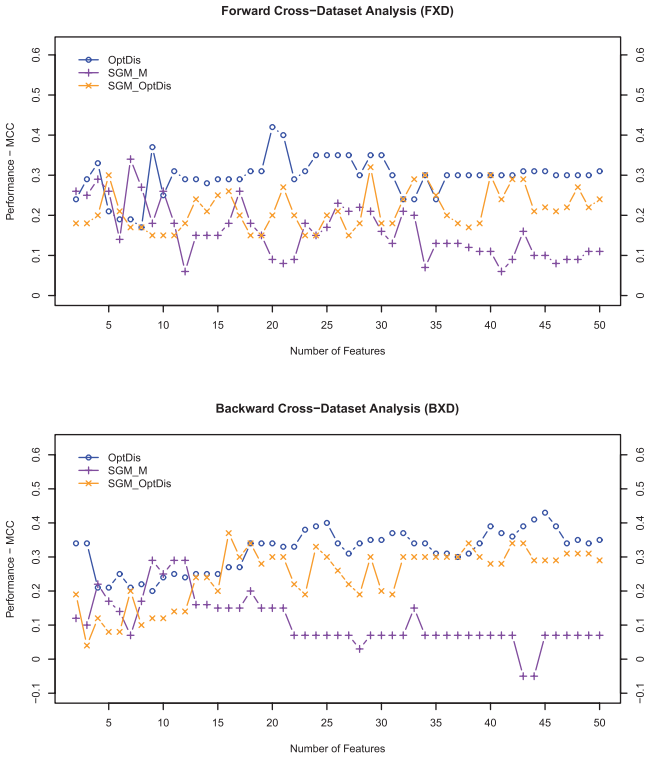

Fig. 5. Line graphs show the MCCs for different predictive models: our approach (OptDis), the model that constitutes genes from top x ($1 \leq x \leq 50$) subnetworks from OptDis (SGM_OptDis) and another model that ranks genes by *t*-test and matches the number of genes from top x ($1 \leq x \leq 50$) subnetworks from OptDis (SGM_M).

3.4 Role of predictive markers in drug response

Gene function analysis: we hypothesized that the set of 39 genes (O39) common between the two T50 SN signatures trained on different cohorts may be important to the activity of TFAC therapy. Some of their biological functions are listed in Table 1. About half are implicated in apoptosis, suggesting that changes in strengths of pro-apoptotic and anti-apoptotic signals can induce resistance to chemotherapy. There are also genes involved in DNA repair, which is expected given many of the anticancer drugs within TFAC therapy induce DNA damage (i.e. cyclophosphamide by cross-linking DNA strands).

Some of the 39 genes have specific functions related to mechanism of individual TFAC drugs. Paclitaxel is a mitotic inhibitor that stabilizes microtubule activity during mitosis and induces cell death. While paclitaxel is known to act on beta-tubulin, some studies (Kavallaris, 2010) have also shown association between the actin and tubulin cytoskeleton in drug response, and suggest that regulation of actin cytoskeleton can induce sensitivity to mitotic-inhibitors. From our O39 list, the EVL, RET and CST3 genes have regulatory roles in the organization and assembly of actin filaments.

Fluouracil's primary anticancer activity blocks DNA replication by suppressing thymidyate synthetase activity and depleting thymidine (Longley *et al.*, 2003). *In vitro* studies have shown that AR and IGF2, from our O39 list, can increase incorporation of thymidine, which acts in antagonist to thymidyate synthetase

suppression, to allow DNA synthesis through the actions of thymidine kinase (Pedram *et al.*, 2007; Yang *et al.*, 1996).

Doxorubicin is an anthracycline antibiotic that intercalates with DNA and causes double-stranded breaks to induce cell apoptosis or disruption in mitosis (Minotti *et al.*, 2004; Munro *et al.*, 2010). SMAD3 from our list has been observed to affect BRCA1-dependent double-stranded DNA break repair in breast cancer cell lines and thus potentially may contribute to differential response to doxorubicin (Dubrovska, 2005).

Signalling pathway analysis: finally, we also compared subnetwork and single gene markers based on their insights into the mechanisms underlying drug response. We derived the T50 SN, T50 SG, and Tx SG from the combined cohort of 230 patients and used the Ingenuity Pathway Analysis software (IPA; Ingenuity© Systems, www.ingenuity.com) to identify significant pathway associations. Interestingly, several signaling pathways associated with chemotherapy response were identified for SN markers, whereas no significantly enriched pathways were found for the T50 and T111 SG markers (Fig. 6). A closer examination of the top associated pathways suggests response to TFAC treatment is affected by the cross-talk between tumor subtype specific mechanisms and pathways regulating apoptosis. Chemotherapy response in breast cancer have been observed to be subtype-specific (Sorlie, 2006), with ER+ tumors exhibiting much higher response rates to taxane-based therapies than ER− tumors (Farmer, 2009; Liedtke, 2008; Popovici, 2010). Therefore, it was expected to find that the predictive subnetwork signature was strongly enriched for genes activating the estrogen receptor (ER) signaling pathway. For the same reason, we also observe an enrichment for the androgen receptor (AR) signaling pathway. With nearly all ER+ tumors and few ER− tumors showing AR expression (Niemeier *et al.*, 2010), it is likely that AR-based subnetworks serve as good predictive markers of TFAC treatment based on their association with ER status. Based on the enriched IPA pathways associated with response, we speculate that the differential response between subtypes may be attributed to differential regulation of apoptosis. Experimental studies have shown that expression of ERα selectively inhibits paclitaxel-induced apoptosis through modulation of glucocorticoid receptor activity (Sui *et al.*, 2007).

Other response-associated pathways may also contribute to differential response to TFAC treatment. For example, signalling of insulin-like growth factor has known functions in cancer proliferation and inhibition of apoptosis, and has been experimentally implicated in chemotherapy resistance (Benini,

Table 1. Table of enriched molecular and cellular functions related to drug response of overlapping gene set (O39)

Enriched terms	Gene symbols	*P*-value
Apoptosis	AR, EP300, ESR1, GADD45G, IGF2, IGF1R, IGFBP4, IL6ST, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, TSC2	1.27E-06
DNA synthesis	AR, ESR1, IGF2, IGFBP4, IL6ST, MDM2, SHC1, SRC	1.74E-06
Actin filament organization	EVL, CST3, RET, SRC, TSC2	7.16E-03
DNA repair	GADD45G, MDM2, RARA, SMAD3	1.89E-02

The *P*-values are adjusted using Benjamini–Hochberg method.

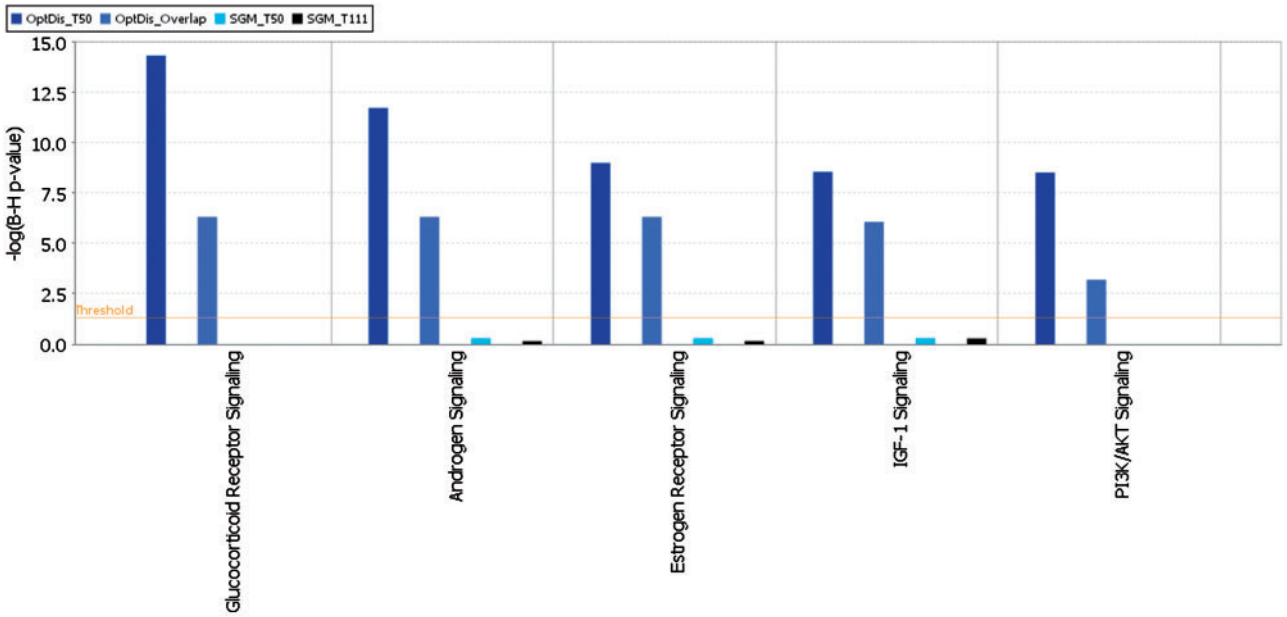

Fig. 6. Signaling pathways associated with TFAC response, ranked by enrichmentin the T50 SN derived from our OptDis method. We also compare the enrichment of those pathways in the genes from T50 SN (dark blue), O39 (light blue), T50 SG (cyan), and T111 SG markers (black). Significantly enriched pathways have Benjamini-Hochberg corrected p-values above threshold of 0.05 (dotted line).

2001; Dunn *et al.*, 1997; Gooch *et al.*, 1999). The PI3K/AKT pathway can also increase resistance to taxane-based therapies through downstream anti-apoptotic effectors BCL-2 and BCL-XL (McGrogan *et al.*, 2008). Experiments have shown that tumors with increased phosphorylated BCL-2 expression have increased sensitivity to paclitaxel compared with tumors with reduced expression (Shitashige *et al.*, 2001).

We measured the reproducibility of these pathway enrichments by performing IPA pathway analysis on both O39 genes and the T50 SNs derived from the pooled 230 patients using another SN method (GreedyMI). Figure 6 shows that both predictive SN signatures were significantly enriched with the same pathways, which may implicate a strong role for these pathways in response to TFAC treatment.

4 CONCLUSIONS

From our analyses, we derived subnetwork markers from separate cohorts of patients and clearly demonstrated the advantages of using subnetwork markers over single gene markers for the prediction chemotherapy response. The improved reproducibility of subnetwork markers and its relevant insights into the underlying mechanisms of drug resistance or sensitivity suggest that they may serve as better clinical predictors of drug response.

Funding: IGTC Mathematical Biology Training Program and Bioinformatics for Combating Infectious Diseases fellowships (to P.D.), CIHR Bioinformatics Training Program (to K.W.).

Conflict of Interest: none declared.

REFERENCES

Alon,N. *et al*. (1995) Color-coding. *J. ACM*, **42**, 844–856.

Alon,N. *et al*. (2008) Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, **24** (Suppl. 1), i241–i249.

Baldi,P. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Benini,S. (2001) Inhibition of insulin-like growth factor I receptor increases the antitumor activity of doxorubicin and vincristine against Ewing's sarcoma cells. *Clin. Cancer Res.*, **7**, 1790–1797.

Bonnefoi,H. *et al*. (2009) Predictive signatures for chemotherapy sensitivity in breast cancer: are they ready for use in the clinic? *Eur. J. Cancer*, **45**, 1733–1743.

Bruckner,S. *et al*. (2010) Topology-free querying of protein interaction networks. In *J. Comput. Biol.*, **17**, 237–252.

BW,M. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta.*, **405**, 442–451.

Chowdhury,S.A. and Koyutürk,M. (2010) Identification of coordinately dysregulated subnetworks in complex phenotypes. In *Pacific Symposium on Biocomputing*, pp. 133–144.

Chowdhury,S.A. (2010) Subnetwork state functions define dysregulated subnetworks in cancer. *J. Comput. Biol.*, **18**, 263–281.

Chuang,H.Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Cleator,S. *et al*. (2006) Gene expression patterns for doxorubicin (Adriamycin) and cyclophosphamide (cytoxan) (AC) response and resistance. Breast Cancer Res. Treat., **95**, 229–233.

Dao,P. *et al*. (2009) Quantifying systemic evolutionary changes by color coding confidence-scored ppi networks. In *9th International Workshop on Algorithms in Bioinformatics*, pp. 37–48.

Dao,P. *et al*. (2010) Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics*, **26**.

Dittrich,M.T. *et al*. (2008) Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**, i223–i231.

Dost,B. *et al*. (2008) Qnet: A tool for querying protein interaction networks. *J. Comput. Biol.*, **15**, 913–925.

Dubrovska,A. (2005) TGFbeta1/Smad3 counteracts BRCA1-dependent repair of DNA damage. *Oncogene*, **24**, 2289–2297.

Dunn,S.E. *et al*. (1997) Insulin-like growth factor 1 (IGF-1) alters drug sensitivity of HBL100 human breast cancer cells by inhibition of apoptosis induced by diverse anticancer drugs. *Cancer Res.*, **57**, 2687–2693.

Ein-Dor,L. *et al*. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl Acad. Sci. USA*, **103**, 5923–5928.

Farmer,P. (2009) A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. *Nat. Med.*, **15**, 68–74.

Fortney,K. *et al*. (2010) Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. *Genome Biol.*, **11**, R13.

Gooch,J.L. *et al*. (1999) Insulin-like growth factor (IGF)-I rescues breast cancer cells from chemotherapy-induced cell death–proliferative and anti-apoptotic effects. *Breast Cancer Res. Treat.*, **56**, 1–10.

Hess,K.R. (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.*, **24**, 4236–4244.

Irizarry,R.A. *et al*. (2003) Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.*, **31**.

Jensen,L. (2009) String 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Johnson,W.E. *et al*. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Kavallaris,M. (2010) Microtubules and resistance to tubulin-binding agents. *Nat. Rev. Cancer*, **10**, 194–204.

King,A.D. *et al*. (2004) Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**, 3013–3020.

Lee,J.K. *et al*. (2007) A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc. Natl Acad. Sci. USA*, **104**, 13086–13091.

Liedtke,C. *et al*. (2010) Clinical evaluation of chemotherapy response predictors developed from breast cancer cell lines. *Breast Cancer Res. Treat.*, **121**, 301–309.

Liedtke,C. (2008) Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J. Clin. Oncol.*, **26**, 1275–1281.

Longley,D.B. *et al*. (2003) 5-fluorouracil: mechanisms of action and clinical strategies. *Nat. Rev. Cancer*, **3**, 330–338.

McGrogan,B.T. *et al*. (2008) Taxanes, microtubules and chemoresistant breast cancer. *Biochim. Biophys. Acta*, **1785**, 96–132.

Minotti,G. *et al*. (2004) Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacol. Rev.*, **56**, 185–229.

Munro,A.F. *et al*. (2010) Targeting anthracyclines in early breast cancer: new candidate predictive biomarkers emerge. *Oncogene*, **29**, 5231–5240.

Niemeier,L.A. *et al*. (2010) Androgen receptor in breast cancer: expression in estrogen receptor-positive tumors and in estrogen receptor-negative tumors with apocrine differentiation. *Mod. Pathol.*, **23**, 205–212.

Pedram,A. *et al*. (2007) A conserved mechanism for steroid receptor translocation to the plasma membrane. *J. Biol. Chem.*, **282**, 22278–22288.

Popovici,V. (2010) Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.*, **12**, R5.

Prasad,T.S. *et al*. (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.*, **577**, 67–79.

Qiu,Y. *et al*. (2009) Identifying differentially expressed pathways via a mixed integer linear programming model. **3**, 475–486.

Scott,J. *et al*. (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol.*, **13**, 133–144.

Shi,L. (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–838.

Shitashige,M. *et al*. (2001) Dissociation of Bax from a Bcl-2/Bax heterodimer triggered by phosphorylation of serine 70 of Bcl-2. *J. Biochem.*, **130**, 741–748.

Shlomi,T. *et al*. (2006) Qpath: a method for querying pathways in a protein–protein interaction network. *BMC Bioinformatics*, **7**, 199.

Sorlie,T. (2006) Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics*, **7**, 127.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.

Su,J. *et al*. (2010) Identification of diagnostic subnetwork markers for cancer in human protein–protein interaction network. *BMC Bioinformatics*, **11** (Suppl. 6), S8.

Sui,M. *et al*. (2007) Estrogen receptor alpha mediates breast cancer cell resistance to paclitaxel through inhibition of apoptotic cell death. *Cancer Res.*, **67**, 5337–5344.

Ulitsky,I. *et al*. (2008) Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In *5th Annual International Conference on Research in Computational Molecular Biology*, pp. 347–359.

van't Veer,L.J. and Bernards,R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**, 564–570.

Wu,Z. *et al*. (2010) A systems biology approach to identify effective cocktail drugs. *BMC Syst. Biol.*, **4** (Suppl. 2), S7.

Yang,C.Q. *et al*. (1996) The expression and characterization of human recombinant proinsulin-like growth factor II and a mutant that is defective in the O-glycosylation of its E domain. *Endocrinology*, **137**, 2766–2773.