Sparse multitask regression for identifying common mechanism of response to therapeutic targets

Kai Zhang, Joe W. Gray and Bahram Parvin*

ABSTRACT

Motivation: Molecular association of phenotypic responses is an important step in hypothesis generation and for initiating design of new experiments. Current practices for associating gene expression data with multidimensional phenotypic data are typically performed one-to-one, i.e. each gene is examined independently with a phenotypic index and tested with one stress condition at a time, i.e. different perturbations are analyzed separately. As a result, the complex coordination among the genes responsible for a phenotypic profile is potentially lost. More importantly, univariate analysis can potentially hide new insights into common mechanism of response. Results: In this article, we propose a sparse, multitask regression model together with co-clustering analysis to explore the intrinsic grouping in associating the gene expression with phenotypic signatures. The global structure of association is captured by learning an intrinsic template that is shared among experimental conditions, with local perturbations introduced to integrate effects of therapeutic agents. We demonstrate the performance of our approach on both synthetic and experimental data. Synthetic data reveal that the multitask regression has a superior reduction in the regression error when compared with traditional $L_1$- and $L_2$-regularized regression. On the other hand, experiments with cell cycle inhibitors over a panel of 14 breast cancer cell lines demonstrate the relevance of the computed molecular predictors with the cell cycle machinery, as well as the identification of hidden variables that are not captured by the baseline regression analysis. Accordingly, the system has identified GLCA2 as a hidden transcript and as a common mechanism of response for two therapeutic agents of CI-1040 and Iressa, which are currently in clinical use.

Contact: b_parvin@lbl.gov

1 INTRODUCTION

Genome-wide association studies of expression and phenotypic data are becoming a routine methodology for identifying potential biomarkers. While the literature is rich with supervised or unsupervised clustering of genomic information, methods for studying the relationships between genomic and phenotypic data remain relatively limited. Existing association methods are typically based on the univariate correlation analysis, which either correlates a single gene to the resultant phenotype(s) or vice versa. This is known as the gene- and phenotype-based approaches, respectively (Dryja, 1997). More recently, (Yi et al., 2008) quantized large number of transcript data through clustering, and associated them with physiological responses or clinical metadata. In contrast, another group of researchers have taken a new direction by first clustering morphometric data and then associating with the transcript data (Han et al., 2010). However, in both cases, correlation is based on the independent, pairwise univariate analysis. Pairwise univariate correlation analysis can quickly provide important association information, as well as candidates for further screening. However, it treats the genes and the phenotypes as independent and isolated units, therefore the underlying interacting relationships between the units might be lost. It is well-known that some transcripts act as regulatory nodes, driving other transcripts in a coordinated manner to determine the phenotypic profile. Additionally, incubation with each therapeutic reagent simultaneously interferes with a subset of genes. Here, we hypothesized that simultaneous incorporation of genome-wide expression data coupled with phenotypic data computed from multiple perturbation conditions, each targeting a different molecular region, can elucidate a common mechanism of response that may be hidden otherwise. In fact, perturbation and molecular diversity of the model system have shown to be capable of reducing the samples needed for biological inference, thus enhancing robustness of biological conclusion (Ideker et al., 2001; Sachs et al., 2005; Tegnér et al., 2003). Thus, we ask the following questions. How can traditional univariate associations be modeled simultaneously and in the absence of a correlation threshold? How can the inherent sparsity of association be formalized within an optimization framework? How can one compensate for the lack of replicates due to the high experimental cost associated with gene expression profiling? To address these issues, we have developed an integrated platform that simultaneously and systematically takes into account an ensemble of gene and phenotypic signatures. Such an enterprise must incorporate an experimental design with sufficient degree of molecular diversity for increased computational robustness. In this context, molecular diversity is achieved by using a panel of breast cancer cell lines that are well-characterized and readily available through American Type Culture Collection.

Our computational framework consists of two major steps. First, a vector-valued, multitask regression formulation is adopted to model the relationships between transcripts and phenotypes under multiple experimental conditions. In particular, the regression coefficients are factorized into two parts. One part is a shared template that suggests a common mechanism of action under various treatments. The second part is related to the perturbation that is induced locally in the transcript network under individual perturbation. The regression has to be sparse, because only a subset of genes is typically involved in a specific phenotypic response. Sparsity is enforced through $L_1$-norm regularization, which inherently removes outliers and irrelevant associations. The end result is a sparse regression matrix that captures intrinsic properties of gene-phenotype association. This matrix is reordered for improved visualization of the gene-phenotype grouping, where the reordering aims at an optimum permutation of rows and columns of the regression matrix such that

*To whom correspondence should be addressed.
the underlying saliency becomes apparent. In this context, reordering reveals dominant association between subsets of genes (with the similar expression profile) and subset of phenotypic indices (with the similar measurements).

We have demonstrated the efficacy of our method with synthetic and experimental data, where the main purpose of synthetic data is to profile the robustness and precision of the proposed method. Experimental data consist of baseline gene expression data for a panel of breast cancer cell lines, which are associated with cell-cycle inhibitor data. The proposed method can be used as a complementary tool besides baseline regression techniques, to provide a richer and a more promising list of candidate molecular predictors for further biological verifications.

Section 2 presents our computational model and detailed optimization procedures. Section 3 provides results on synthetic and experimental data. Section 4 concludes with a discussion on the molecular predictors and system performance.

2 MODELS

2.1 Description of basic computational models

In this section, we introduce our basic computational models for exploring the associations between genes and phenotypic responses. To reduce excessive costs associated with the collection of gene expression data, we assumed that the gene expression were collected under a baseline (unperturbed) condition, as denoted by $X_0 \in \mathbb{R}^{C \times N}$. Here, $C$ is the number of cell lines and $N$ is the number of genes. On the phenotypic side, assume that we obtained measurements $Y_d \in \mathbb{R}^{C \times M}$ for $d = 0, 1, 2, \ldots, D$, where $M$ is the number of phenotypic features, $d = 0$ denotes the controlled, baseline condition and $d = 1, 2, \ldots, D$ corresponds to the drug-perturbed conditions. We used the linear regression model to measure the dependency between genes and phenotypes, as illustrated in Figure 1. The design matrix $X_0$ was mapped to the phenotype responses $Y_d \in \mathbb{R}^{C \times M}$ via a reconstructing matrix $T_d \in \mathbb{R}^{C \times M}$, as

$$ X_0 T_d \rightarrow Y_d. \quad (1) $$

The coefficient matrices $T_d$'s reflect the dependency (or correlation) between the genes and the phenotypes of interest, i.e. its $ij$-th entry is the weight associated with the $i$-th gene in reconstructing the $j$-th feature in the phenotypic profile under the $d$-th condition.

There are a number of complexities in estimating $T$. These complexities originate from low sample size, high dimensionality of the data and coupling between different perturbation conditions. However, majority of the transcript data can be considered as noisy background, as it believed that only a subset of genes are involved in each specific cellular process. To address these issues, we propose a sparse, regularized multitask regression framework with co-clustering. The novelty of our method involves: (i) leveraging the locality of the molecular interactions as a result of treatment with therapeutic agents, and modeling multiple experiments simultaneously; (ii) coupling it with a $L_1$-regularized solution that enforces sparsity and simultaneously compensates for small sample size; and (iii) grouping associations with co-clustering.

First, a multitask regression framework is used to model the molecular interactions under multiple conditions in a systematic way. The Multitask learning (Caruana, 1997; Lee et al., 2007; Xiong et al., 2007) is aimed at information sharing among learners from a set of different but related tasks, with the hope to boost the overall performance. In this context, regression (1) under each experimental condition is deemed as a task. As phenotypic profiles arise from the original gene regulatory network and its local perturbation, we can assume that phenotypic responses are triggered by different experimental conditions are lying on the same low-dimensional space, i.e.

$$ T_d = T \cdot P_d \quad \text{for } d = 0, 1, 2, \ldots, D. \quad (2) $$

In other words, task relatedness is enforced by requiring that $T_d$’s associated with each task are local perturbations of a shared subspace $T$. Here, $T \in \mathbb{R}^{N \times K}$ represents the shared structure (related to the gene regulatory network), $P_d \in \mathbb{R}^{K \times M}$ compensates for the perturbation of different experimental conditions and $K$ is the dimension of the latent space in which the phenotypic responses are supposed to reside. In our formulation, $K$ is set to be equal to $M$ for practical reasons, and $P_d$'s are diagonal matrices. The actual structure of $P_d$ is an open problem at this point, and it is possible that a non-diagonal matrix can produce a better reconstruction result.

The structure of $P_d$ and the choice of $K$ is one of the topics for our continued research. Nevertheless, the shared template matrix $T$ has the potential to summarize association descriptor between $N$ genes and $M$ phenotypes. An advantage of decomposing the $T_d$ matrices is a significant reduction in the number of variables for estimation.

Second, the $L_1$ regularization technique is used to mathematically guarantee the robustness of the system against irrelevant genes. The $L_1$ regularization typically leads to sparse learning models, and has been independently discovered in several research areas such as regression shrinkage and variable selection (Tibshirani, 1996), basis pursuit (Donoho et al., 2001), compressive sensing (Donoho, 2006) and feature vector machine (Li et al., 2005). By penalizing the $L_1$-norm of the variables, part of the regression coefficients will be driven to zero with the level of sparsity controlled by the strength of regularization. This is a desirable property considering the highly localized functionalities of genes as they relate to specific phenotypic signatures.

By combining the multitask learning frame with the $L_1$ regularization, we established sparse multitask regression as follows:

$$ \min_{T \in \mathbb{R}^{C \times M}, P_d \in \mathbb{R}^{K \times M}} f = \sum_{d=0}^{D} ||X_0 T P_d - Y_d||_F^2 + \lambda \|T\|_1, \quad \text{s.t. } \|P_d\|_F = 1, \text{ for } d = 1, 2, \ldots, D. \quad (3) $$

Fig. 1. The linear regression model used to compute the sparse association between baseline gene expression data and phenotypic responses.
Sparse multitask regression

Formulation (3) is a vector-valued regression with intrinsic $T$ and perturbation-specific $P_d$'s. It can be solved by an alternating optimization strategy, i.e., iteratively fixing $P_d$'s and solving $T$, and then fixing $T$ and solving $P_d$'s. We will show that both $T$ and $P_d$'s subproblems are convex. Thus a locally optimal solution of the problem (3) can always be guaranteed. In the following, we present details on the alternating optimization (Parts I, II and III) and the co-clustering procedure (Part IV).

(I) Fix $(P_d)_{d=0}^D$ and solve $T$: We will show that when $P_d$'s are fixed, $T$ can be solved through quadratic programming. First, use the operator vec $(\bullet)$: $\mathbb{R}^{p \times q} \rightarrow \mathbb{R}^{pq 	imes 1}$ to define the mapping $T$ which transforms a $p \times q$ matrix into a $pq \times 1$ vector via concatenating the columns in the matrix, and let $\text{vec}(\bullet)$ be the inverse mapping. Let $t = \text{vec}(T) \in \mathbb{R}^{MN \times 1}$. Then define a 3D matrix $A_d \in \mathbb{R}^{C \times M \times MN}$ for $d = 0, 1, 2, \ldots, D$, such that

$$A_d(i,j,:) = \text{vec}(X_d(i,:)^T \cdot P_d(j,:)^T) \quad (4)$$

Here, $X_d(i,:)$ is the $i$-th row in $X_d$, $P_d(j,:)$ the $j$-th column in $P_d$, and each $(i,j)$-pair locates an $MN \times 1$ vector denoted by $A_d(i,j,:)$. Now, computing $T$ is equivalent to the following quadratic program

$$\min_{t \in \mathbb{R}^{MN \times 1}} t^T Q t - 2b^T t + \lambda \|t\|_1 \quad (5)$$

where $Q = \sum_{d=0}^D \sum_{i,j=1}^{C,M} A_d(i,j,:)^T A_d(i,j,:) \quad (6)$

and $b = \sum_{d=0}^D \sum_{i,j=1}^{C,M} Y_d(i,j)^T A_d(i,j,:) \quad (7)$

It can be easily verified that the residual term

$$\sum_{d=0}^D \|X_d TP_d - Y_d\|_F^2$$

in (3) is identical to $t^T Q t - 2b^T t$ up to a constant that is independent of the optimization variables. Note that the Hessian of the above quadratic programming problem is positive semi-definite: for any $x \in \mathbb{R}^{MN \times 1}$ we have

$$x^T Q x = \sum_{d=0}^D \sum_{i,j=1}^{C,M} A_d(i,j,:)^T A_d(i,j,:)^T x \geq 0.$$
As \( Q \) is a positive semi-definite (PSD) matrix, there exists the spectrum of \( Q \) decays rapidly, with only the top 48 eigenvalues being strictly non-zero, thus substantiating the low-rank nature of the \( Q \) matrix. As a result, the Hessian matrix can be represented by the 'low-rank approximation' to alleviate prohibitive computational difficulties. We therefore pursued an approximate solution by using the low-rank decomposition (8) allows us to rewrite the \( L_1 \)- regularized quadratic programming problem (5) into a standard least square problem with \( L_1 \) regularization.

\[
\min_{t \in \mathbb{R}^{m \times 1}} \frac{1}{2} \|t - q\|^2 + \lambda \|t\|_1. \tag{9}
\]

Here, \( q \in \mathbb{R}^{m \times 1} \) can be determined by expanding the quadratic term in (9), comparing it with (3) and requiring \( L_q \approx -h \). Formulation of (9) is a good approximation to the original problem (5) and it has been widely examined in statistics, optimization and machine learning. We use the 11-ls solver (Kim et al., 2007) for large-scale \( L_1 \) regularized least square problems, which are based on the truncated Newton interior-point method. Empirically, it can solve large sparse problems with a million variables with high accuracy in a few tens of minutes on a modern desktop computer.

(I) Fix \( T \) and solve \( P_T d_q \). By fixing \( T \), entries of \( P_T \)'s can be computed using simple scalar equations. Let the \( i \)-th column of the matrix \( X \) be denoted by \( X(i,:) \) and the \( i \)-th column in \( Y_d \) be \( Y_d(i,:) \). It’s easy to verify that the \( i \)-th diagonal entry in \( P_T \) can be solved easily as \( P_T(i,i) = X(i,:)'X(i,:)'Y_d(i,:) / \|Y_d(i,:)\|^2 \). To guarantee that \( P_T \)'s all have Norm 1, we will normalize them by \( P_T = P_T / \|P_T\|_F \). This can be deemed as iteratively projecting the solutions on the feasible region \( \|P_T\|_F = 1 \).

Note that rescaling both \( T \) and \( P_T \)'s with \(-1\) does not affect the prediction performance of the multitask regression, but will reverse the signs of associations learned in \( T \). To solve this problem, we require that the signs of the resultant matrix \( T \) should be maximally correlated with those of the standard correlation coefficients on the same set of genes. From a practical standpoint, because \( P_T \)'s are initialized with identity matrices, we have always observed that they continue to be PSD during the optimization procedure. Empirically, our method converges rapidly in about 5 to 10 iterations on our current datasets.

(III) Initialization and parameter selection: By fixing one of the two groups of variables, \( F \) or \( P_T \)'s \((d = 1,2,\ldots,D)\), the other can be computed. Here, we choose to initialize \( P_T \)'s as identity matrices for \( d = 1,2,\ldots,D \). Note that initialization of the \( T \)'s is usually much easier than that of \( T \), where degrees of freedom are \( M^2D \) and \( MN \), respectively. We used leave-one-out cross-validation to choose the hyperparameter \( \lambda \) since the sample size is very small. This involves selecting one sample as a testing sample and the rest as training. We repeated this process for each sample and computed the averaged predictor error on the testing sample at each grid point \( \lambda \in \{10^{-3},10^{-2},10^{-1},1,10\} \).

(IV) Co-clustering: Template \( T \) is an intrinsic regression coefficient matrix linking the gene expression and phenotypic signature under the multiple conditions studied: the \( j \)-th entry signifies the strength of the relationship between the \( i \)-th gene and the \( j \)-th phenotype. To reveal the clustered structure in these associations, we use co-clustering to permute the rows and columns of \( T \), so that the underlying saliency becomes apparent and can be visualized. We have adopted the bipartite spectral clustering (Dhillon, 2001) for simultaneously clustering the genes and phenotypes. Bipartite spectral clustering uses a bipartite graph where vertices are divided into two types, each from one dimension of the given contingency table \( T \). In our case they are genes and phenotypes, denoted by \( G \) and \( P \), respectively, and the number of vertices will be \( M+N \). The edge weights are determined by \( W_{ij} = t(X(i,j)) \). \( v_i \in G \) and \( v_j \in P \). In other words, edges only exist between a gene vertex and a phenotype vertex. By applying
sparse multitask regression

3 RESULTS

Our proposed method has been tested with both synthetic and experimental data. The synthetic data is used for method validation and profiling against other known techniques. Our studies with experimental data identified molecular predictors of cell cycle data from baseline gene expression data.

3.1 Evaluation with synthetic data

In the synthetic case: (i) a data matrix $X \in \mathbb{R}^{50 \times 300}$ was created from the Gaussian distribution; (ii) a sparse intrinsic template $T \in \mathbb{R}^{300 \times 5}$ with 50 non-zero rows and a small set of randomly generated perturbation matrices $P_d \in \mathbb{R}^{300 \times 5}$ were created for each $d = 1, 2, \ldots, D$ task; and (iii) the responses (e.g. target values) were then determined by $Y_d = X_d T P_d + \epsilon$, where $\epsilon$ is the noise term. We examined how well the system recovers $T P_d$'s, and compared the proposed method with (i) independent $L_1$-regularized regression, and (ii) independent $L_2$-regularized regression, also known as regularized least squares (RLS). First, we set $D = 10$ and selected one of the tasks to visualize the regression qualities against the competing methods. Reconstruction results are shown in Figure 4. Notice that the $L_2$ and $L_2$ regressions (Fig. 4c and d) 'contaminated' the true regression coefficients. In practical association analysis, this can lead to a number of false predictions. In contrast, multitask regression (Fig. 4b) reliably recovered the regression coefficients. Second, we varied $D$ from 1 to 50 and quantified the average per-task-error for each of the three methods, as shown in Figure 5. It is clear that the error in multitask regression decreases monotonically with the number of tasks, while the errors in pure $L_1$ and $L_2$ regressions remain stationary. Although this experiment demonstrates an improved error profile for multitask learning, we have not yet designed a synthetic experiment that maintains a correlation between transcripts.

3.2 Experimental design and quantification of biological endpoints

We applied our method to a set of publicly available gene expression data for a panel of breast cancer cell lines collected with Affymetrix HG-U133A (Neve et al., 2006). We used the following 14 cell lines: MCF7, HCC1806, HCC1954, HCC1428, A55651, MDAMB415, SUM185PE, ZR75B, MCF7, MDAMB361, LY2, T47D, MDAMB436, MDAMB468 and ZR75F. From the original $N = 22,215$ probe sets, we chose 5706 by removing those with a variance of $<0.3$. This is slightly above the noise level of the Affymetrix U133 platform. Notice that the gene expression data were collected under baseline (e.g. unperturbed) condition. Our main

| Independent L1-reg | Independent L2-reg | Multitask L1-reg |
|-------------------|-------------------|------------------|
| -1.0 | -0.8 | -0.6 | -0.4 | -0.2 | 0.0 | 0.2 | 0.4 |
| 10 | 20 | 30 | 40 | 50 | 60 | 70 |

Fig. 4. Reconstruction of the regression coefficient matrix indicates that multitask learning is more accurate when compared with $L_1$- and $L_2$-regularized regressions. $T_d$ is a 300-by-5 matrix and each column is represented by a unique color. (a) Ground-truth solution, (b) Multitask regression, (c) standard $L_1$ regression and (d) regularized least square regression.

Fig. 5. Multitask learning has an improved error rate profile as the number of tasks is increased.

hurdle has been the prohibitive cost of collecting necessary data (e.g. three conditions, 14 lines, and at least three biological replicates). Thus, we assumed that perturbed expression data would be linearly predictable from the control data.

Cell cycle data where collected for cells exposed to three conditions: control condition (e.g. DMSO solvent alone), the MEK inhibitor CI1040 and the tyrosine kinase inhibitor Iressa. Both these inhibitors induce cell cycle arrest, but through different mechanisms. Each cell line was plated in triplicate and incubated for 48h with CI1040 and Iressa at 5.6 and 4.0 $\mu$M, respectively. Subsequently, samples were fixed and stained with Hoechst and BrdU, and 25 fields of view were imaged using the Celomics high-throughput system. These images were uploaded into the BioSig imaging bioinformatics system (Parvin et al., 2003), and then analyzed for their morphometric and BrdU incorporation on a cell-by-cell basis (Raman et al., 2007; Wen et al., 2009). Figure 6 shows a sample of images that have been registered with the BioSig and one segmented image. Each segmented nucleus is represented using a multidimensional feature (Han et al., 2010) and stored in the database. In our experiment, the pertinent features are total BrdU and
Fig. 6. (a) Biological images are registered with BioSig and (b) each nucleus is segmented to quantify total DNA and BrdU incorporation on a cell-by-cell basis.

Fig. 7. By aggregating total DNA and BrdU, on a cell-by-cell basis for all images in each well, the percentages of cells in G1, S, and G2 phase are quantified.

Fig. 8. Percentage of each cell line being arrested in G1 phase with DMSO, CI1040, and Iressa treatment conditions.

3.3 Evaluation with therapeutic agents

First, we examined associations of gene expression and cell cycle data using independent $L_1$-regularized regression that learns the regressing coefficients $T_{d,j}$’s separately for each experimental condition. The results enabled us to contrast traditional $L_1$ regression with multitask learning. Predicted results are shown in Figure 9, where each subfigure corresponds to the regression matrix $T_{d,j}$ under one condition. Here, zero rows in the regression matrix were removed, and the rows and columns of $T_{d,j}$’s have been reordered by the co-clustering procedure. The positive and negative association between each gene-phenotype pair is encoded by green and red blocks, respectively. Second, we applied the proposed multitask regression to learn a common template of correlation between genes and cell cycle data for the two inhibitors (e.g. CI1040 and Iressa), as shown in Figure 10. Again, we assumed that each therapeutic reagent would perturb a small molecular region in the cell cycle progression. In this experiment, both CI1040 and Iressa induced cell cycle arrest by targeting different molecular moieties. However, if there is a common mechanism of action, then we would like to infer that. We observed that the genes identified by multitask regression (Fig. 10) contained subset of genes that were identified separately by independent $L_1$ regression, shown in Figures 9b and c.
Sparse multitask regression

Fig. 9. The regression matrices (a) $T_0$ (DMSO), (b) $T_1$ (C11040), and (c) $T_2$ (Iressa) learned by the independent regression using 14 cell lines and reordered by co-clustering.

Fig. 10. The intrinsic template $T$ learned by the multitask regression using 14 cell lines and the two drug conditions (C11040 and Iressa) and reordered by co-clustering.

However, there are certain genes that can only be predicted through the multitask regression. These are hidden markers that are relevant to the effect of the therapeutic reagent and provide potential new hypothesis for further studies. The total computation time on a modern desktop computer is approximately 6500 s.

4 DISCUSSION

Our experiments with synthetic data have clearly demonstrated that multitask learning offers the following advantages over independent $L_1$ regression: (i) regression is less noisy; (ii) regression error is reduced as a function of the number of tasks; and (iii) hidden variables are revealed since traditional $L_1$ regression can push non-zero coefficients to zero and vice versa. Therefore, the bulk of the discussion in this section is devoted to the experimental data by focusing on a few important genes and their independent analysis through Ingenuity Pathway Analysis (IPA) and Pathway Studio.

(I) CLCA2 is a hidden variable that has been identified through multitask regression and is shown to be negatively associated with the S phase. We hypothesized that CLCA2 is a common mechanism of response for inhibitors C11040 and Iressa. This gene is known to be downregulated in breast cancer cell lines. In addition to being a p53 client (Gruber and Pauli, 1999), its knockdown leads to increased invasiveness (Walia et al., 2009), and it is epigenetically regulated (Li et al., 2004). It is also a tumor suppressor gene that may be a potential target for therapy. It is likely that CLCA2 acts as a common molecular switch to inhibit DNA synthesis and initiate apoptosis as a result of treatment with either therapeutic agent. Therefore, it not only serves as a therapeutic target, but can also be used in combination with other therapeutic targets used today for improved lethality.

(II) NLRP2 is regulated by NFκB and is shown to be expressed in MDA-MB-436 and MCF-7 (Bruey et al., 2004) breast cancer cell lines. This particular gene appears in both independent and multitask regression. Furthermore, the Gene Ontology annotation indicates that NLRP2 is in involved in caspase activities and apoptosis. We hypothesized that strong G1 arrest and complementary negative correlation with cells being in S is the result of treatment with the therapeutic agent. This particular gene is reflected in multitask regression and independent regression analysis corresponding to C11040 and Iressa. It is also a potential common mechanism of response for further analysis.

(III) CDKN2A (also known as p16) expression is positively associated with G1 arrest in normal cells and tissues, but is negatively associated with the S phase in our analysis of the human
normally regulated by JUN and FOS (Takahashi et al., 1998), whose gene products together constitute the AP1 transcription factor. AP1 drives the expression of a number of genes that are necessary for cell cycle progression. The relationships between these protein–protein interactions are shown in Figure 11. This gene appears in multitask and one of the independent regression analysis.

Three common regulators that have been inferred from a subset of genes associated with the S phase.

CA2 is an example of the gene that is reported by both independent association of gene expression data with CH1040 (Fig. 9b) and the multitask regression analysis (Fig. 10). CA2 is ordinarily involved in differentiation and apoptosis, overexpressed in MCF7 and MDA-MB-231 and negatively correlated with the S phase in the drug-treated cells. SiRNA-mediated interference with this gene. The aberrations result in continued proliferation in the drug-treated cells.

Finally, we performed an independent analysis by using Ingenuity Pathway Analysis and Pathway Studio, scientific software that helps researchers more effectively search, explore, visualize, and analyze biological and chemical findings related to genes, proteins and small molecules. We selected the set of genes that was correlated with the drug, negative regulation of cell proliferation, negative regulation of apoptosis, overexpressed in MCF7 and MDA-MB-231 and negatively correlated with the S phase in the drug-treated cells. SiRNA-mediated interference with human CA2 gene expression has been shown to decrease survival of MDA-MB-231 cell lines (Mallery et al., 2005).

In summary, multitask learning has the potential to summarize a vast amount of data, compute biologically relevant markers and identify hidden variables that traditional regressors may fail to capture. Although the technique is currently applied for integration of gene expression data with cell cycle data, it can also be used for other integrative biology applications.

ACKNOWLEDGEMENTS

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

REFERENCES

Baker, C.T.H. (1997) The Numerical Treatment of Integral Equations. Clarendon Press: Oxford.

REFERENCES

Donoho, D.L. (2006) Compressed sensing. 

Dhillon, I.S. (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY.

Ding, C.H.Q. (2003) Unsupervised feature selection via two-way ordering in gene expression analysis. Bioinformatics, 19, 1259–1266.

Dryja, T.F. (1997) Gene-based approach to human gene-phenotype correlations. Proc. Natl Acad. Sci. USA, 94, 12117–12121.

Gauthier, M.L. et al. (2007) Abrogated response to cellular stress identifies DCIS associated with subsequent tumor events and defines basal-like breast tumors. IEEE Trans. Inf. Theory, 12, 479–491.
Sparse multitask regression

Golub, G.H. and Loan, C.F.V. (1996) Matrix Computations 3rd Edition (Johns Hopkins Studies in Mathematical Sciences). The Johns Hopkins University, Baltimore, MD.
Gruber, A.D. and Pauli, B.U. (1999) Tumorigenicity of human breast cancer is associated with loss of the ca2+-activated chloride channel CLCA2. *Cancer Res.*, 59, 5488–5491.
Han, J. et al. (2010) Multidimensional profiling of cell surface proteins and nuclear marker. *IEEE Trans. Comput. Biol. and Bioinform.*, 7, 80–90.
Han, J. et al. (2010) Molecular predictors of 3D morphogenesis by breast cancer cell lines in 3D culture. *PLoS Comput. Biol.*, 6, e1000864.
Hartigan, J.A. (1972) Direct clustering of a data matrix. *J. Am. Stat. Assoc.*, 67, 123–129.
Idler, T. et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292, 929–934.
Kim, J.I. et al. (2007) An interior-point method for large-scale 1-regularized least squares. *IEEE J. Sel. Top. Signal Process.*, 1, 606–617.
Kluger, Y. et al. (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, 13, 703–716.
Lee, S.I. et al. (2007) Learning a meta-level prior for feature relevance from multiple related tasks. In *Proceedings of the 24th International conference on Machine learning*, ACM, New York, NY.
Li, F. et al. (2005) From lasso regression to feature vector machine. In *Advances in Neural Information Processing Systems 18*, MIT Press, Cambridge, MA.
Li, X. et al. (2004) CLCA2 tumour suppressor gene in 1p31 is epigenetically regulated in breast cancer. *Oncogene*, 23, 1474–1480.
Mallory, J.C. et al. (2005) A novel group of genes regulates susceptibility to antineoplastic drugs in highly tumorigenic breast cancer. *SIAM Rev.*, 48, 1747–1756.
Nevins, J.R. et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell*, 10, 315–327.
Parvin, B. et al. (2003) Biosig: an imaging bioinformatics system for phenotypic studies. *IEEE Trans. Syst. Man Cybern.*, B, 33, 814–824.
Raman, X. et al. (2007) Geometric approach segmentation and protein localization in cell cultured assays. *J. Microsc.*, 225, 22–30.
Sanada, K. et al. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 525–529.
Takahashi, H. et al. (1994) Structure and transcriptional regulation of the human cystatin a gene. *J. Bio. Chem.*, 273, 17375–17380.
Tanay, A. et al. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18, 136–144.
Tegnér, J. et al. (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl Acad. Sci. USA*, 100, 5944–5949.
Tibshirani, R. (1996) Regression shrinkage and selection via the ILASSO. *J. R. Stat. Soc., Series B*, 58, 267–288.
Walik, V. et al. (2009) IdCLCA2 is a p53-inducible inhibitor of breast cancer cell proliferation. *Cancer Res.*, 69, 6624–6632.
Wen, Q. et al. (2009) A Delunay triangulation approach for segmenting clumps of nuclei. In *Proceedings of the IEEE International Symposium on Biomedical Imaging: from nano to macro*, Boston, MA, 9–12.
Xiong, Y. et al. (2007) Probabilistic joint feature selection for multi-task learning. In *SIAM International Conference on Data Mining*, Minneapolis, MN.
Yang, X. et al. (2009) Heterogeneous multitask learning with joint sparsity constraints,. In *Proceeding of the 23rd Neural Information Processing Systems*, MIT press, Cambridge, MA, 2151–2159.
Ye, S.G. et al. (2008) Response projected clustering for direct association with physiological and clinical response data. *BMC Bioinformatics*, 9, 76.