BIOINFORMATICS

Vol. 26 ISMB 2010, pages i97–i105
doi:10.1093/bioinformatics/btq181

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

Kai Zhang, Joe W. Gray and Bahram Parvin*

Life Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

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 (i) performed one-to-one, i.e. each gene is examined independently with a phenotypic index and (ii) 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 CLCA2 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 $\mathbf{X}_0 \in \mathbb{R}^{C\times N}$. Here, C is the number of cell lines and N is the number of genes. On the phenotypic side, assume that we obtained measurements $\mathbf{Y}_d \in R^{C\times M}$'s for $d=0,1,2,....,D$, where M is the number of phenotypic features, $d=0$ denotes the controlled, baseline condition and $d=1,2,...,D$ 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 $\mathbf{X}_0$ was mapped to the phenotype responses $\mathbf{Y}_d \in \mathbb{R}^{C\times M}$ via a regressing matrix $\mathbf{T}_d \in \mathbb{R}^{N\times M}$, as

$$\mathbf{X}_0\mathbf{T}_d \rightarrow \mathbf{Y}_d. \tag{1}$$

The coefficient matrices $\mathbf{T}_d$'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 treatments 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.

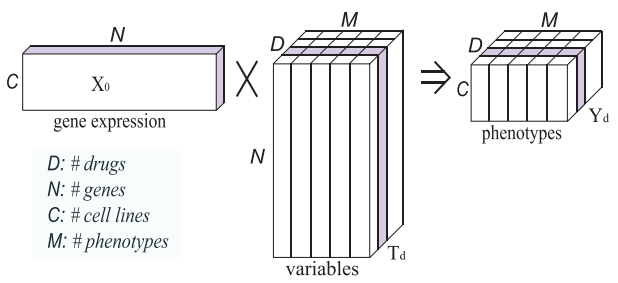

Fig. 1. The linear regression model used to compute the *sparse* association between baseline gene expression data and phenotypic responses.

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.

$$\mathbf{T}_d = \mathbf{T}\cdot\mathbf{P}_d \text{ for } d=0,1,2,...,D. \tag{2}$$

In other words, task relatedness is enforced by requiring that $\mathbf{T}_d$'s associated with each task are local perturbations of a shared subspace $\mathbf{T}$. Here, $\mathbf{T}\in\mathbb{R}^{N\times K}$ represents the shared structure (related to the gene regulatory network), $\mathbf{P}_d \in \mathbb{R}^{K\times M}$ 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 $\mathbf{P}_d$'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 $\mathbf{T}$ has the potential to summarize association descriptor between N genes and M phenotypes. An advantage of decomposing the $\mathbf{T}_d$ 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:

$$\min_{\substack{\mathbf{T}\in\mathbb{R}^{N\times M}\\ \mathbf{P}_d\in\mathbb{R}^{M\times M}}} f=\sum_{d=0}^{D}\|\mathbf{X}_0\mathbf{T}\mathbf{P}_d-\mathbf{Y}_d\|_F^2+\lambda\|\mathbf{T}\|_1. \tag{3}$$

$$\text{s.t.} \qquad \|\mathbf{P}_d\|_F=1, \text{ for } d=1,2,...,D.$$

Here, $\|\cdot\|_F$ is the matrix Frobenius norm and $\|\cdot\|_1$ is the matrix L_1-norm. The first term enforces a fit between the gene expression

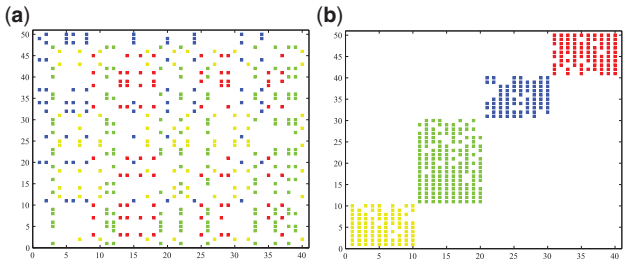

Fig. 2. The co-clustering procedure transforms a randomly displayed association table (**a**) of 50 genes and 40 phenotypes to a organized partition (**b**).

and the phenotypic signature under each condition, while the second term enforces sparsity on the shared template $\mathbf{T}$. The constraints $\|\mathbf{P}_d\|=1$ are used to prevent trivial solutions (i.e. $\mathbf{T}$ approaches zero and $\mathbf{P}_d$'s approach infinity). Alternatively, this can be achieved by penalizing $\|\mathbf{P}_d\|_F$ with a extra regularization parameter. More recently, a heterogeneous multitask learning framework that considers both continuous (regression) and discrete (classification) variables was successfully used to discover genetic markers that jointly influence multiple correlated traits (Yang *et al.*, 2009). In comparison, our method considers pure regression setting only, where the phenotypic measurements are continuous.

Formulation (3) allows us to obtain condition-specific regression matrices $\mathbf{T}_d$'s based on a common template $\mathbf{T}$. Note that for each $\mathbf{T}_d$, its non-zero rows signify important genes under the d-th condition. Therefore, template $\mathbf{T}$, which is shared among multiple $\mathbf{T}_d$'s, defines a combined list of genes that are important to the phenotypes studied under these conditions. In other words, $\mathbf{T}$ is an integrated association descriptor that summarizes correlating relations between genes and phenotypes under multiple conditions; and we want to read out useful structures (such as the grouped correlation between subsets of genes and subsets of phenotypes) encoded in $\mathbf{T}$. To achieve this goal, we performed co-clustering analysis (Hartigan, 1972) on $\mathbf{T}$. Co-clustering analysis has been used to find clusters in various tabulated data such as the co-occurrence of documents/words (Dhillon, 2001), or the expression of genes under various conditions (Ding, 2003; Kluger *et al.*, 2003; Tanay *et al.*, 2002), by simultaneously grouping rows and columns of the association table. However, it has rarely been applied to interpret associations between genes and phenotypes, where the association table is not directly available from raw data but instead has to be learned. In fact, co-clustering can reorganize regression coefficients in a perceptually meaningful manner to bring more insights into our analysis. This is illustrated by synthetic data, as shown in Figure 2. For example, assume we have learned an association table of 50 rows (e.g. genes) and 40 columns (e.g. phenotypes) where it is difficult to observe any meaningful structures. However, if we permute the rows and columns of the table by co-clustering (Dhillon, 2001), we will discover four dominant correlation groups, as shown in the Figure 2B. Such a grouping can be regarded as a distinctive 'watermark' of the gene–phenotypic association. Furthermore, rows (genes) grouped into the same block are more likely to participate together in affecting corresponding columns (phenotype responses).

In summary, the sparse multitask regression has three advantages: (i) it allows us to reduce the number of variables from $O(MND)$ to $O(NM+DM^2)$; (ii) the sparsity of $\mathbf{T}$ easily transfers to those of $\mathbf{T}_d$'s due to the simple linear relation $\mathbf{T}_d=\mathbf{T}\cdot\mathbf{P}_d$; and (iii) as we shall see, the template matrix $\mathbf{T}$ is a platform from which explorative analysis can be carried out in identifying important, grouped correspondences between genes and phenotypic signatures.

2.2 Optimization procedures

Formulation (3) is a vector-valued regression with intrinsic $\mathbf{T}$ and perturbation-specific $\mathbf{P}_d$'s. It can be solved by an alternating optimization strategy, i.e. iteratively fixing $\mathbf{P}_d$'s and solving $\mathbf{T}$, and then fixing $\mathbf{T}$ and solving $\mathbf{P}_d$'s. We will show that both $\mathbf{T}$ and $\mathbf{P}_d$'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 $\{\mathbf{P}_d\}_{d=0}^D$ and solve $\mathbf{T}$: We will show that when $\mathbf{P}_d$'s are fixed, $\mathbf{T}$ can be solved through quadratic programming. First, use the operator $\mathrm{vec}(\cdot):\mathbb{R}^{p\times q}\to\mathbb{R}^{pq\times 1}$ to denote the mapping that transforms a $p\times q$ matrix into a $pq\times 1$ vector via concatenating the columns in the matrix, and let $\mathrm{ivec}(\cdot)$ be the inverse mapping. Let $\mathbf{t}=\mathrm{vec}(\mathbf{T})\in\mathbb{R}^{MN\times 1}$. Then define a 3D matrix $\mathbf{A}_d\in\mathbb{R}^{C\times M\times MN}$ for $d=0,1,2,...,D$, such that

$$\mathbf{A}_d(i,j,:)=\mathrm{vec}\left(\mathbf{X}_0(i,:)^\top\cdot\mathbf{P}_d(:,j)^\top\right). \tag{4}$$

Here, $\mathbf{X}_0(i,:)$ is the i-th row in $\mathbf{X}_0$, $\mathbf{P}_d(:,j)$ the j-th column in $\mathbf{P}_d$ and each (i,j)-pair locates an $MN\times 1$ vector denoted by $\mathbf{A}_d(i,j,:)$. Now, computing $\mathbf{T}$ is equivalent to the following quadratic program

$$\min_{\mathbf{t}\in\mathbb{R}^{MN\times 1}}\quad \mathbf{t}^\top\mathbf{Q}\mathbf{t}-2\mathbf{b}^\top\mathbf{t}+\lambda\|\mathbf{t}\|_1 \tag{5}$$

$$\text{where}\quad \mathbf{Q}=\sum_{d=0}^{D}\sum_{i=1}^{C}\sum_{j=1}^{M}\mathbf{A}_d(i,j,:)\mathbf{A}_d(i,j,:)^\top \tag{6}$$

$$\mathbf{b}=\sum_{d=0}^{D}\sum_{i=1}^{C}\sum_{j=1}^{M}\mathbf{Y}_d(i,j)\mathbf{A}_d(i,j,:). \tag{7}$$

It can be easily verified that the residual term $\sum_{d=0}^{D}\|\mathbf{X}_0\mathbf{T}\mathbf{P}_d-\mathbf{Y}_d\|_F^2$ in (3) is identical to $\mathbf{t}^\top\mathbf{Q}\mathbf{t}-2\mathbf{b}^\top\mathbf{t}$ 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 $\mathbf{x}\in\mathbb{R}^{MN\times 1}$ we have

$$\mathbf{x}^\top\mathbf{Q}\mathbf{x}=\sum_{d=0}^{D}\sum_{i=1}^{C}\sum_{j=1}^{M}\mathbf{x}^\top\mathbf{A}_d(i,j,:)\mathbf{A}_d(i,j,:)^\top\mathbf{x}$$

$$=\sum_{d=0}^{D}\sum_{i=1}^{C}\sum_{j=1}^{M}\left(\mathbf{A}_d(i,j,:)^\top\mathbf{x}\right)^2\geq 0.$$

On the other hand, the L_1 regularization term $\lambda\|\mathbf{t}\|_1$ is a convex function. Therefore, the problem is convex, and there exists a unique, globally optimal solution for the subproblem (5).

The main computational barrier is that the Hessian matrix $\mathbf{Q}$ is MN-by-MN, which can be very large and does not fit in a modern desktop computer. However, this matrix is symmetric, positive-definite Hessian matrix $\mathbf{Q}$ and has very low rank in practice, i.e. its eigen-spectrum decays very quickly to zero. This is shown in Figure 3, where we chose $N=1210$ genes and $M=3$ phenotypes to construct the matrix $\mathbf{Q}$ (6) with size 3630×3630. It is clear that

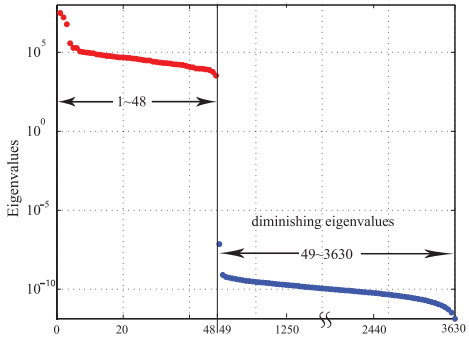

Fig. 3. Spectrum of a 3630×3630 matrix $\mathbf{Q}$, computed from our experimental data, indicates that only the largest 48 eigenvalues are strictly positive and the rest are insignificant. The spectrum clearly reflects the low-rank nature of the matrix $\mathbf{Q}$ and the feasibility of low-rank approximation.

the spectrum of $\mathbf{Q}$ decays rapidly, with only the top 48 eigenvalues being strictly non-zero, thus substantiating the low-rank nature of the $\mathbf{Q}$ matrix. As a result, the Hessian matrix can be represented by the 'low-rank approximation' to alleviate prohibitive computational requirements. To do this, we searched for a rank-R matrix $\mathbf{L}$ that best represents the $\mathbf{Q}$ matrix in a least square sense, $\min_{\mathbf{L}\in\mathbb{R}^{MN\times R}}\|\mathbf{Q}-\mathbf{L}\mathbf{L}^\top\|_F^2$, where $R\ll NM$, $\mathbf{L}\in\mathbb{R}^{MN\times R}$ is a rectangular matrix with low row-rank and $\mathbf{L}\mathbf{L}^\top$ is called the rank-R approximation of $\mathbf{Q}$. This approximation $\mathbf{Q}\approx\mathbf{L}\mathbf{L}'$ dramatically reduces memory usage from $O(N^2M^2)$ to $O(NMR)$.

Mathematically, the optimal rank-R matrix $\mathbf{L}$ is given by the eigenvectors of $\mathbf{Q}$ (Golub and Loan, 1996), which is computationally expensive. We therefore pursued an approximate solution by adopting the sampling-based low-rank approximation scheme, known as the Nyström method, which originated from the numerical treatment of integral equations of the second type (Baker, 1997). The basic idea of the Nyström method is to randomly sample R columns from the $\mathbf{Q}$ matrix, which, due to its symmetry, also corresponds to R rows. Let E and E' denote the sampled columns and its transpose, respectively, where $E\in\mathbb{R}^{MN\times R}$. Let $W\in\mathbb{R}^{R\times R}$ be the intersection of the selected rows and columns. Then $\mathbf{Q}$ can be decomposed as $\mathbf{Q}\approx EW^{-1}E'$. In our specific context, $\mathbf{Q}$ is represented as the sum of multiple outer products (6). By utilizing this property, E and W can be computed efficiently as follows:

$$E(p,q)=\sum_{d=0}^{D}\sum_{i=1}^{C}\sum_{j=1}^{M}\mathbf{A}_d(i,j,p)\mathbf{A}_d(i,j,q),$$

$$W=E(I,I),\ 1\le p\le MN,\ q\in I,$$

where $I=\{1,2,...,MN\}^R$ is the index of selected columns. Given W and E, the low-rank approximation of $\mathbf{Q}$ is then expressed as

$$\mathbf{Q}\approx\mathbf{L}\mathbf{L}^\top,\ \text{where}\ \mathbf{L}=EW^{-\frac{1}{2}}. \tag{8}$$

As W is a positive semi-definite (PSD) matrix, there exists theoretically a real square root of W. In practice, we could encounter diminishing eigenvalues. A robust way is to first perform the eigenvalue decomposition $W=U\Lambda U^\top$, remove those diminishing eigenvalues and then let $W^{\frac{1}{2}}=U\Lambda^{\frac{1}{2}}U^\top$.

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),

$$\min_{\mathbf{t}\in\mathbb{R}^{MN\times1}}\ \left\|\mathbf{L}'\mathbf{t}-\mathbf{q}\right\|^2+\lambda\|\mathbf{t}\|_1. \tag{9}$$

Here, $\mathbf{q}\in\mathbb{R}^{R\times1}$ can be determined by expanding the quadratic term in (9), comparing it with (3) and requiring $\mathbf{L}'\mathbf{q}=\mathbf{b}$. 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 l1-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.

(II) Fix $\mathbf{T}$ and solve $\{\mathbf{P}_d\}_{d=1}^D$: By fixing $\mathbf{T}$, entries of $\mathbf{P}_d$'s can be computed using simple scalar equations. Let the i-th column of the matrix $\mathbf{X}\cdot\mathbf{T}$ be denoted by $\mathbf{XT}(:,i)$ and the i-th column in $\mathbf{Y}_d$ be $\mathbf{Y}_d(:,i)$. It's easy to verify that the i-th diagonal entry in $\mathbf{P}_d$ can be solved easily as $\mathbf{P}_d(i,i)=\mathbf{XT}(:,i)^\top\mathbf{XT}(:,i)/\|\mathbf{Y}_d(:,i)\|_2^2$. To guarantee that $\mathbf{P}_d$'s all have Norm 1, we will normalize them by $\mathbf{P}_d=\mathbf{P}_d/\|\mathbf{P}_d\|_F$. This can be deemed as iteratively projecting the solutions on the feasible region $\|\mathbf{P}_d\|_F=1$.

Note that rescaling both $\mathbf{T}$ and $\mathbf{P}_d$'s with -1 does not affect the prediction performance of the multitask regression, but will reverse the signs of associations learned in $\mathbf{T}$. To solve this problem, we require that the signs of the resultant matrix $\mathbf{T}$ should be maximally correlated with those of the standard correlation coefficients on the same set of genes. From a practical standpoint, because $\mathbf{P}_d$'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, T or $\mathbf{P}_d$'s ($d=1,2,...,D$), the other can be computed. Here, we choose to initialize P_d's as identity matrices for $d=1,2,...,D$. Note that initialization of the $\mathbf{T}_d$'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 λ 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 $\mathbf{T}$ is an intrinsic regression coefficient matrix linking the gene expression and phenotypic signature under the multiple conditions studied: the ij-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 used co-clustering to permute the rows and columns of $\mathbf{T}$, 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 ($\mathbf{T}$). In our case they are genes and phenotypes, denoted by $\mathcal{G}$ and $\mathcal{P}$, respectively, and the number of vertices will be $M+N$. The edge weights are determined by $W_{ij}=\begin{cases}|\mathbf{T}(i,j)| & v_i\in\mathcal{G},v_j\in\mathcal{P},\\ 0 & v_i,v_j\in\mathcal{G}\ \text{or}\ v_i,v_j\in\mathcal{P}.\end{cases}$ In other words, edges only exist between a gene vertex and a phenotype vertex. By applying

spectral clustering on this bipartite graph, simultaneous groupings on gene and phenotype vertices can be computed. Mathematically, we need to compute the singular value decomposition of the degree-normalized association matrix, $\mathbf{S}=\mathbf{D}_l^{-\frac{1}{2}}\mathbf{T}\mathbf{D}_r^{-\frac{1}{2}}$, where $\mathbf{D}_l$ is an $N\times N$ diagonal degree matrix whose i-th entry is the summation of the i-th row in $\mathbf{T}$, and $\mathbf{D}_r$ is a $M\times M$ diagonal degree matrix whose i-th diagonal entry is the summation of the i-th column of $\mathbf{T}$. Interestingly, the left and right singular vectors of $\mathbf{S}$ (corresponding to the second largest singular value) not only provide a partitioning of the rows and columns of $\mathbf{T}$, but also provide a natural ordering (embedding) of the required row and column permutations.

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 $\mathbf{X}_0\in\mathbb{R}^{50\times 300}$ was created from the Gaussian distribution; (ii) a sparse intrinsic template $\mathbf{T}\in\mathbb{R}^{300\times 5}$ with 50 non-zero rows and a small set of randomly generated perturbation matrices $\mathbf{P}_d\in\mathbb{R}^{5\times 5}$ were created for each $d=1,2,...,D$ task; and (iii) the responses (e.g. target values) were then determined by $\mathbf{Y}_d=\mathbf{X}_0\mathbf{T}\mathbf{P}_d+\epsilon$, where ϵ is the noise term. We examined how well the system recovers $\mathbf{T}_d$'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_1 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: MCF12A, HCC38, HCC1428, AU5650, MDAMB415, SUM185PE, ZR75B, MCF7, MDAMB361, LY2, T47D, MDAMB436, MDAMB468 and ZR751. 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

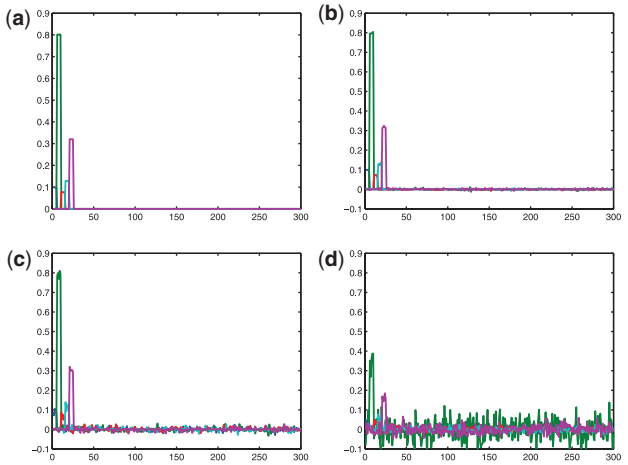

Fig. 4. Reconstruction of the regression coefficient matrix indicates that multitask learning is more accurate when compared with L_1- and L_2-regularized regressions. $\mathbf{T}_d$ is a 300-by-5 matrix and each column is represented by a unique color. (**a**) Ground-truth solution, (b) Multitask regression,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

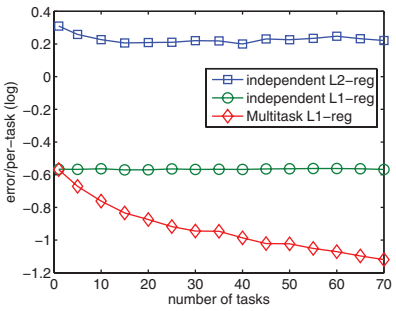

Fig. 5. Multitask learning has an improved error rate profile as the number of tasks is increased.

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 48 h with CI1040 and Iressa at 5.6 and 4.0 μM, 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

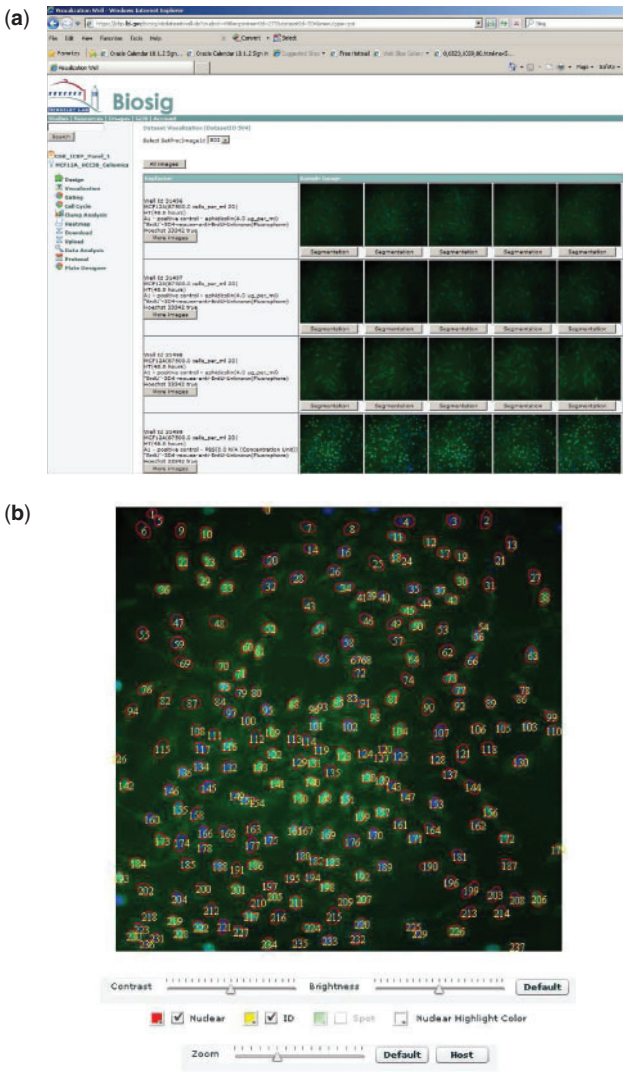

Fig. 6. (**a**) Biological images are registered with BioSig and (**b**) each nucleus is segmented to quantify total DNA and BrdU incorporation on a cell-by-cell basis.

DNA content on a cell-by-cell basis. By aggregating these features, within each well, percentages of cells being G_1, S and G_2 Phase can be quantified as a function of their treatment, as shown in Figure 7. The main advantage of microscopy for evaluating cell cycle arrest is a significant reduction in the number of required cells. Finally, outliers were removed. Summary results are shown in Figure 8.

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 $\mathbf{T}_d$'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 $\mathbf{T}_d$ under one condition. Here, zero rows in the regression matrix were removed, and the rows and columns of $\mathbf{T}_d$'s have been reordered by the co-clustering procedure. The positive and negative association

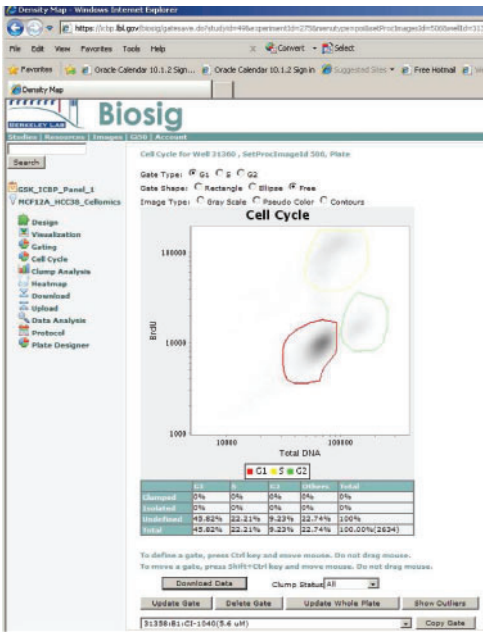

Fig. 7. By aggregating total DNA and BrdU, on a cell-by-cell basis for all images in each well, the percentages of cells in G_1, S, and G_2 phase are quantified.

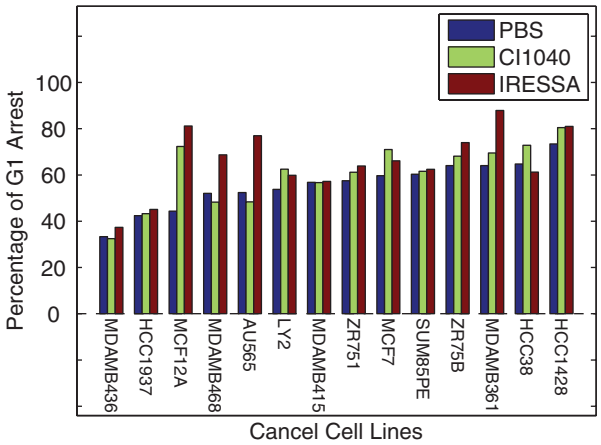

Fig. 8. Percentage of each cell line being arrested in G_1 phase with DMSO, CI1040, and Iressa treatment conditions.

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.

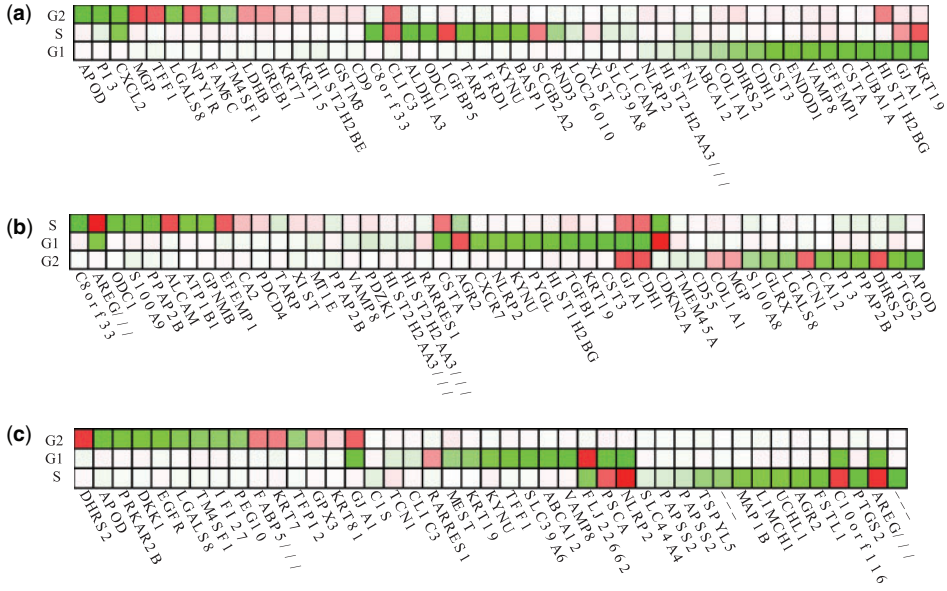

Fig. 9. The regression matrices (**a**) $\mathbf{T}_0$, (DMSO) (**b**)$\mathbf{T}_1$ (CI1040), and (**c**)$\mathbf{T}_2$ (Iressa) learned by the independent regression using 14 cell lines and reordered by co-clustering.

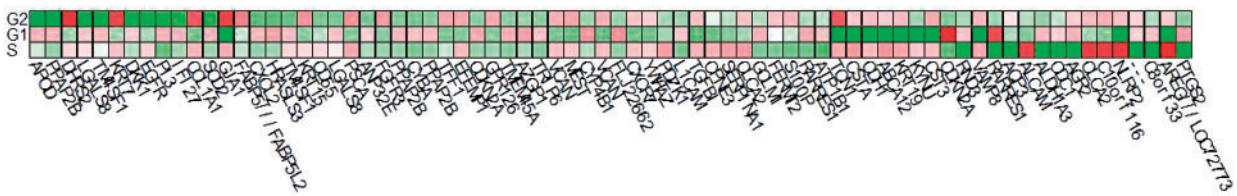

Fig. 10. The intrinsic template **T** learned by the multitask regression using 14 cell lines and the two drug conditions (CI1040 and Iressa) and reordered by co-clustering.

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 CI1040 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 CI1040 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

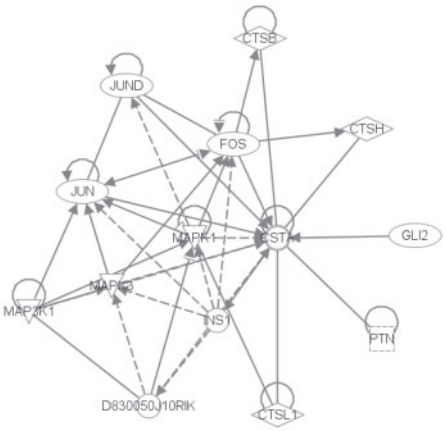

Fig. 11. Interaction of CSTA with JUN and FOS curated through IPA.

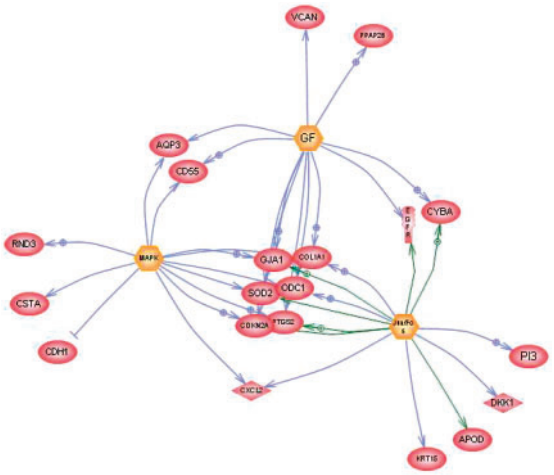

Fig. 12. Three common regulators that have been inferred from a subset of genes associated with the S phase.

breast tumor cell lines (in both the independent regression of Fig. 9b and the multitask regression of Fig. 10). This discrepancy is likely explained by the fact that most of the malignant cell lines in the panel have aberrations in downstream effectors of the product of this gene. The aberrations result in continued proliferation in the presence of p16 expression that ordinarily would yield cell cycle arrest and senescence (Gauthier *et al*., 2007).

(IV) CSTA is involved in apoptosis and differentiation, and is 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.

(V) CA2 is an example of the gene that is reported by both independent association of gene expression data with CI1040 (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 human CA2 gene expression has been shown to decrease survival of MDA-MB-231 cell lines (Mallory *et al*. , 2005).

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 S phase, and uploaded them into IPA and Pathway Studio. The IPA analysis indicated that this group of genes is largely involved in (i) cell cycle and signaling *networks* and (ii) cancer. The net result is a more substantial support for gene-by-gene analysis. Similar results have been obtained from Pathway Studio, which provides gene set enrichment analysis (GSEA) and identifies common regulators with the user-defined number of neighbors. Gene enrichment analysis revealed that predicted gene groups are involved in response to toxin, drug, negative regulation of cell proliferation, negative regulation of peptidase activity where S phase is one of them and apoptosis among top-ranked groups. Furthermore, a number of common regulators with high *P*-values were also inferred that are associated with the cell cycle machinery. Figure 12 shows three regulators of MAPK, Jun/Fos, and GF, and their target entities.

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.

Funding: U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (contract DE-AC02-05CH11231); the National Institutes of Health (grants U54 CA112970 and CA58207).

Conflict of Interest: none declared.

REFERENCES

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

Bruey,J.M. *et al*. (2004) Pan1/nalp2/pypaf2, an inducible inflammatory mediator that regulates nf-kappab and caspase-1 activation in macrophages. *J. Bio. Chem.*, **279**, 51897–51907.

Caruana,R. (1997) Multitask learning. *Mach. Learn.*, **28**, 41–75.

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.

Donoho,D.L. (2006) Compressed sensing. *IEEE Trans. Inf. Theory*, **52**, 1289–1306.

Donoho,D.L. *et al*. (2001) Atomic decomposition by basis pursuit. *SIAM Rev.*, **43**, 129–159.

Dryja,T.P. (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.

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 profilng 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 Computat. Biol.*, **6**, e1000684.

Hartigan,J.A. (1972) Direct clustering of a data matrix. *J. Am. Stat. Assoc.*, **67**, 123–129.

Ideker,T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Kim,S.J. *et al*. (2007) An interior-point method for large-scale l1-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 anti-neoplastic drugs in highly tumorigenic breast cancer. *SIAM Rev.*, **468**, 1747–1756.

Neve,RM. *et al*. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell*, **10**, 515–527.

Parvin,B. *et al*. (2003) Biosig: an imaging bioinformatics system for phenotypic studies. *IEEE Trans. Syst. Man Cybern.*, B, **33**, 814–824.

Raman,S. *et al*. (2007) Geometric approach segmentation and protein localization in cell cultured assays. *J. Microsc*, 225, 22–30.

Sachs,K. *et al*. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Takahashi,H. *et al*. (1998) 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 lLASSO. *J. R. Stat. Soc., Series B*, **58**, 267–288.

Walia,V. *et al*. (2009) hCLCA2 is a p53-inducible inhibitor of breast cancer cell proliferation. *Cancer Res.*, **16**, 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,T. *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.

Yi,S.G. *et al*. (2008) Response projected clustering for direct association with physiological and clinical response dat. *BMC Bioinformatics*, **9**, 76.