BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 14 2012, pages 1911–1918
doi:10.1093/bioinformatics/bts285

Data and text mining

Advance Access publication May 10, 2012

iFad: an integrative factor analysis model for drug-pathway association inference[†]

Haisu Ma[1] and Hongyu Zhao[2,*]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511 and [2]Division of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Pathway-based drug discovery considers the therapeutic effects of compounds in the global physiological environment. This approach has been gaining popularity in recent years because the target pathways and mechanism of action for many compounds are still unknown, and there are also some unexpected off-target effects. Therefore, the inference of drug-pathway associations is a crucial step to fully realize the potential of system-based pharmacological research. Transcriptome data offer valuable information on drug-pathway targets because the pathway activities may be reflected through gene expression levels. Hence, it is of great interest to jointly analyze the drug sensitivity and gene expression data from the same set of samples to investigate the gene-pathway–drug-pathway associations.

Results: We have developed iFad, a Bayesian sparse factor analysis model to jointly analyze the paired gene expression and drug sensitivity datasets measured across the same panel of samples. The model enables direct incorporation of prior knowledge regarding gene-pathway and/or drug-pathway associations to aid the discovery of new association relationships. We use a collapsed Gibbs sampling algorithm for inference. Satisfactory performance of the proposed model was found for both simulated datasets and real data collected on the NCI-60 cell lines. Our results suggest that iFad is a promising approach for the identification of drug targets. This model also provides a general statistical framework for pathway-based integrative analysis of other types of -omics data.

Availability: The R package ‘iFad’ and real NCI-60 dataset used are available at http://bioinformatics.med.yale.edu/group/.

Contact: hongyu.zhao@yale.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on February 15, 2012; revised on April 6, 2012; accepted on May 5, 2012

1 INTRODUCTION

Identification of drug targets, the gene products that bind to specific therapeutic molecules, is essential for understanding drugs’ action mechanism and possible side effects, and for maximizing treatment efficacy and minimizing drug toxicity. Traditional pharmaceutical research and development process is deeply rooted in the ‘one target-one drug’ mindset, which tries to interfere the pathological process through blocking an important molecular player (e.g. a specific enzyme) using a compound. Unfortunately, most drug candidates identified through high-throughput screening based on this philosophy have failed due to either poor efficacy or serious side effects (Schadt *et al*., 2009).

High hope has been placed on using systems biology for drug discovery through the application of novel computational methods to high-throughput genomics and proteomics data. In contrast to the traditional ‘one target-one drug’ perspective that ignores the intricate interaction among genes/proteins, systems biology approaches consider the drug effects in the global physiological environment. They may be more effective in drug discoveries because it has become apparent that most common diseases result from system-level malfunctions rather than problems with individual genes (Pujol *et al*., 2010). In practice, drug combination or poly-pharmacy is more commonly used in recent years to modulate multiple drug targets.

Existing computational methods for drug-target identification can be generally categorized into three classes. The first class of methods uses the classical gene expression profiling strategies to infer drug targets. For example, drug targets can be identified by comparing the mRNA responses in human cell lines induced by drugs with unknown mechanisms with a set of well understood drugs. The connectivity map project (Lamb *et al*., 2006) used a non-parametric rank-based pattern matching strategy based on the Kolmogorov–Smirnov statistic, which requires extensive data mining procedures. Another study (Kutalik *et al*., 2008) proposed a bi-clustering method, named iterative signature algorithm (ISA), to search for ‘co-modules’ representing gene–drug associations.

The second class of methods aims to integrate various types of biological data, including knowledge about bioactive molecules and their protein targets (e.g. sequences, structures and molecular mechanisms) along with the phenotypic effects of drug treatment, to deduce the probability of two proteins being bound by the same ligand (Czodrowski *et al*., 2009; Ecker *et al*., 2008; Nigsch *et al*., 2009). Drug docking studies have also been incorporated to measure the binding probabilities based on 3D structural complementarities (Boyce *et al*., 2009; Irwin *et al*., 2009; Kolb *et al*., 2009; Zavodszky and Kuhn, 2005). One disadvantage of these methods is that they only evaluate the likelihood of presumed drug-target pairs and make no inference of unknown drug targets. Campillos *et al*. (2008) proposed a probabilistic network to calculate the probability of two drugs sharing the same target by comparing their clinical side effects. However, this method provides limited information on the drug action mechanisms.

[*]To whom correspondence should be addressed.

[†]The analysis in this article was performed in ‘Yale University Biomedical High Performance Computing Center’.

The third class of methods analyzes the global patterns of drug–protein interactions (Kuhn *et al*., 2008; Yeh *et al*., 2006; Yildirim *et al*., 2007). It is known that established gene–drug binary associations form a dense network that exhibits local clustering of similar drug types. Such networks can enhance our understanding of the physiological effects of various drugs in terms of the molecular pathways and disease categories involved. However, these ‘guilt-by-association’ methods are too coarse-grained to make quantitative inference about the effect of drugs at the system level.

In this article, we develop a coherent statistical framework to jointly analyze drug sensitivity and gene expression patterns from the same set of cell lines to infer the target pathways of drugs. Much information on these two types of data has been accumulated in the literature (Bussey *et al*., 2006; Ikediobi *et al*., 2006; Shankavaram *et al*., 2007; Sharma *et al*., 2010; Shoemaker, 2006), and if appropriately analyzed, these data may be informative for inferring drug targets. We adopt a latent factor analysis approach, where each latent factor corresponds to the activity of a specific pathway. Note that factor analysis models have been proposed for the reconstruction of gene regulatory networks and the inference of transcription factor activity profiles (Gharib *et al*., 2006; Meng *et al*., 2011; Yeh *et al*., 2009; Yu and Li, 2005). A previous review article (Pournara and Wernisch, 2007) compared the performance of five different factor analysis algorithms. However, these models were developed for the analysis of a single data type, i.e. gene expression data. In contrast, our proposed Bayesian sparse factor analysis model, iFad (integrative factor analysis for drug-pathway association inference), is the first effort to bring together two distinct data types in a unified framework to identify drug-target pathways. We propose and implement a modified collapsed Gibbs sampling algorithm for model inference. Our approach can also easily incorporate known pathway information to infer the ‘many-to-many’ correspondence between the two types of data. Some unique features of our model are (1) joint analysis of distinct data types; (2) a Bayesian framework to integrate prior pathway knowledge and (3) explicit consideration of the sparse nature of the drug-target pathways. Both simulation studies and applications to real NCI-60 datasets show that iFad is a promising approach for drug-target inference.

The rest of the article is organized as follows. We detail the modeling assumptions and statistical inferential procedure in Section 2. Simulations and real data analysis are described in Section 3. We conclude the article in Section 4.

2 METHODS

2.1 Model description

This section describes the statistical framework of our proposed Bayesian sparse factor analysis model, iFad and statistical inference of the model parameters. As discussed above, iFad aims to analyze paired gene expression data and drug sensitivity data generated from the same set of samples. We denote the gene expression dataset by matrix Y_1, with dimension G_1 by J, where G_1 is the number of genes and J is the sample size. The drug sensitivity dataset is denoted by matrix Y_2, with dimension G_2 by J, where G_2 is the number of drugs. Drug sensitivity is usually quantified as the ‘GI_{50}’ values, concentrations required to inhibit growth by 50% (Staunton *et al*., 2001). Matrices Y_1 and Y_2 are normalized (scaled to mean 0 and SD 1 for each gene/drug) before analysis.

iFad links the two matrices through the activity levels of K biological pathways (e.g. KEGG pathways), which are latent factors in our model. The rationale here is that pathway activities influence both gene expression levels and the sensitivity to drugs targeting these in these pathways. We assume that there is some prior knowledge about the gene-pathway and drug-pathway association relationships, represented by two binary matrices L_1 and L_2, with dimensions G_1 by K and G_2 by K, respectively, where $L_1[g,k]=1$ (or $L_2[g,k]=1$) indicates that the gth gene (or drug) is known to be associated with the kth pathway. This information can be retrieved from various pathway databases with different degrees of sensitivity and specificity (Bader *et al*., 2006; Kanehisa *et al*., 2010). iFad assumes that both matrices Y_1 and Y_2 are related to the common underlying pathway activity matrix X (with dimension K by J) through the following linear models:

$$\begin{aligned}
&Y_1=W_1X+\Sigma_1,\Sigma_1\sim N(0,\Psi_1),\Psi_1=\mathrm{diag}\{\tau_{g_1}^{-1}\},\\
&Y_2=W_2X+\Sigma_2,\Sigma_2\sim N(0,\Psi_2),\Psi_2=\mathrm{diag}\{\tau_{g_2}^{-1}\},\\
&X_{k,j}\sim \mathrm{Normal}(0,1),\\
&\tau_{g_1}\sim \mathrm{Gamma}(\alpha_1,\beta_1),g_1=1,2,\dots,G_1,\\
&\tau_{g_2}\sim \mathrm{Gamma}(\alpha_2,\beta_2),g_2=1,2,\dots,G_2.
\end{aligned}$$

Matrices W_1 and W_2 are the factor loading matrices describing the regulatory direction (positive or negative) and strength of the pathway activities on the gene expression levels Y_1 and drug sensitivity Y_2. The latent factor activity matrix X is shared between the two feature spaces, namely gene expression data and drug sensitivity data. Each entry in matrix X is assumed to follow a standard normal distribution. Σ_1 and Σ_2 represent the noise term added to gene expression or drug sensitivity, with mean 0 and diagonal covariance matrices Ψ_1 and Ψ_2. The precision τ_{g1} (for the g_1th gene) and τ_{g_2}(for the g_2th drug) are modeled using a Gamma prior with shape parameters α_1, α_2 and rate parameters β_1, β_2.

In order to use the prior knowledge on the gene-pathway and drug-pathway associations (matrices L_1 and L_2), we use the spike-and-slab mixture prior (West, 2003) for the factor loading matrices W_1 and W_2. Although there exist other forms of sparsity-inducing priors (Pournara and Wernisch, 2007), the spike-and-slab prior has the advantage of easy incorporation of prior information on the connectivity structure of the loading matrix. For both W_1 and W_2, we put the following prior on each entry:

$$\begin{aligned}
&P(W_{g,k})=(1-\pi_{g,k})\delta_0(W_{g,k})+\pi_{g,k}\mathrm{Normal}(W_{g,k}|0,\tau_w^{-1})\\
&\tau_w\sim \mathrm{Gamma}(\alpha_w,\beta_w),
\end{aligned}$$

where δ_0 is the unit point mass at zero (the Dirac delta function) and $\pi_{g,k}$ denotes the prior probability that $W_{g,k}$ is non-zero. If $W_{g,k}$ is non-zero, it is assumed to follow a normal distribution with mean 0 and precision τ_w. The precision τ_w can be either set to a constant or assumed to follow a Gamma prior with parameter (α_w, β_w). Usually, an auxiliary indicator variable $Z_{g,k}$ is used to enable the calculation of posterior probabilities (as Z_1, W_1, π_1, L_1 and Z_2, W_2, π_2, L_2 have very similar formats except the subscript, we just listed the general formula here):

$$P(Z_{g,k}=1)=\pi_{g,k}=\begin{cases}\eta_0, & \text{if } L_{g,k}=0\\ 1-\eta_1, & \text{if } L_{g,k}=1\,.\end{cases}$$

In this way, prior link matrices L_1 and L_2 are used to induce the sparsity structure of the factor loading matrices W_1 and W_2 in a flexible way, with the strength of guidance tuned conveniently by user-specified parameters η_0 and η_1. Under this setting, we can derive the prior probability of different components of the model, as well as the complete joint posterior probability (see Supplementary Materials for details).

It is worth noting that during the simulation studies in Section 3.1, both matrices Z_1 and Z_2 are unknown and are the target of inference. In contrast, for real data analysis (the NCI-60 dataset) in Section 3.2, since prior information about gene-pathway association structure is available and fairly accurate, the major interest lies in the inference of matrix Z_2, the drug-pathway association relationships.

2.2 Inference algorithm

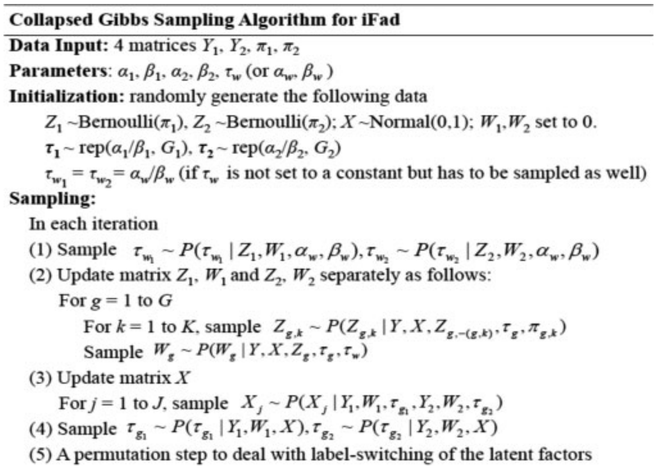
There are many parameters to estimate for iFad. Gibbs sampling is a widely used technique to approximate the joint distribution through re-sampling. However, standard Gibbs sampler may have poor mixing due to

Table 1. Model parameter settings considered in the simulations

Setting	α_g	β_g	α_w	β_w
1	0.7	0.3	0.7	0.3
2	0.7	0.3	σ_w=1	
3	1	0.1	0.7	0.3
4	1	0.1	σ_w=1	
5	1	0.01	0.7	0.3
6	1	0.01	σ_w=1	
7	1	0.005	0.7	0.3
8	1	0.005	σ_w=1	

α_g and β_g are the shape and rate parameters of the Gamma prior put on the precision of the noise term for both matrices Y_1 and Y_2; α_w and β_w are Gamma parameters for the precision of the non-zero elements of the factor loading matrices W_1 and W_2.

dependence between matrices W and Z makes. Therefore, we used a modified collapsed Gibbs sampling algorithm for model inference as outlined below. Detailed derivations of the posterior conditional distributions are provided in Supplementary Materials. At the end of each sampling iteration, we add a local permutation step (Sharp *et al.* 2010 to address the problem of label-switching, which is also described in Supplementary Materials. We have implemented the above algorithm as the R package ‘iFad’, which is publicly available on CRAN.



```
Collapsed Gibbs Sampling Algorithm for iFad
Data Input: 4 matrices $Y_1, Y_2, \pi_1, \pi_2$
Parameters: $\alpha_1, \beta_1, \alpha_2, \beta_2, \tau_w$ (or $\alpha_w, \beta_w$)
Initialization: randomly generate the following data
    $Z_1 \sim$Bernoulli($\pi_1$), $Z_2 \sim$Bernoulli($\pi_2$); $X \sim$Normal(0,1); $W_1, W_2$ set to 0.
    $\tau_1 \sim$ rep($\alpha_1/\beta_1$, $G_1$), $\tau_2 \sim$ rep($\alpha_2/\beta_2$, $G_2$)
    $\tau_{w_1} = \tau_{w_2} = \alpha_w/\beta_w$ (if $\tau_w$ is not set to a constant but has to be sampled as well)
Sampling:
  In each iteration
  (1) Sample $\tau_{w_1} \sim P(\tau_{w_1} | Z_1, W_1, \alpha_w, \beta_w), \tau_{w_2} \sim P(\tau_{w_2} | Z_2, W_2, \alpha_w, \beta_w)$
  (2) Update matrix $Z_1$, $W_1$ and $Z_2$, $W_2$ separately as follows:
      For $g = 1$ to $G$
          For $k = 1$ to $K$, sample $Z_{g,k} \sim P(Z_{g,k} | Y, X, Z_{g,-(g,k)}, \tau_g, \pi_{g,k})$
          Sample $W_g \sim P(W_g | Y, X, Z_g, \tau_g, \tau_w)$
  (3) Update matrix $X$
      For $j = 1$ to $J$, sample $X_j \sim P(X_j | Y_1, W_1, \tau_{g_1}, Y_2, W_2, \tau_{g_2})$
  (4) Sample $\tau_{g_1} \sim P(\tau_{g_1} | Y_1, W_1, X), \tau_{g_2} \sim P(\tau_{g_2} | Y_2, W_2, X)$
  (5) A permutation step to deal with label-switching of the latent factors
```

3 RESULTS

We first tested the performance of iFad using simulated datasets, and then applied the method to real NCI-60 datasets to infer unknown drug-pathway associations.

3.1 Simulation study

In order to assess the performance of our proposed model, we first simulated a series of datasets to investigate the effects of different model parameter settings, as well as the various dataset properties, including sample size and noise level, among other factors.

3.1.1 Data simulation for model parameter selection We tested eight different settings for the Gamma density parameters related to the precision of the gene/drug noise term (τ_{g_1}, τ_{g_2}) and the factor loading matrix (τ_{w_1}, τ_{w_2}), as shown in Table 1.

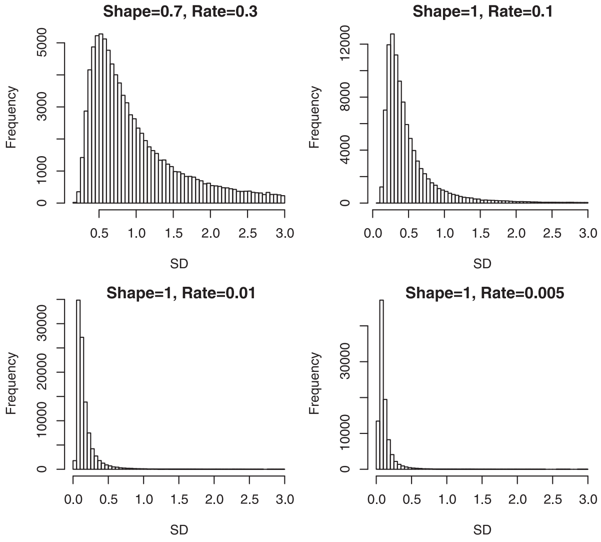


Fig. 1. Histograms of the standard deviations sampled from the Gamma densities with different parameter settings. Note that the Gamma prior is put on the precision τ and SD = 1/sqrt(τ)

Different Gamma parameters represent different prior belief regarding the distribution of the standard deviation for the noise term/the factor loadings. Figure 1 shows the histogram of 10 000 standard deviation values randomly generated from the Gamma density with four parameter combinations tested (only the values smaller than 3 are plotted here). We first simulated two sets of data to compare the eight combinations of parameter settings, with each dataset consisting of four matrices Y_1, Y_2, π_1 and π_2, as shown in Table 2. For both datasets, we used $\alpha_g = 1$, $\beta_g = 0.01$ and $\sigma_w = 1$. For Gibbs sampling, we set the number of iteration to 30 000, with the first half discarded as burn-in period (this was chosen based on the MCMC (Markov chain Monte Carlo) trace plot). To reduce the effect of auto-correlation between adjacent iterations and the data storage burden, we only recorded the Gibbs sampling results every other 10th iteration. We ran five independent chains for Set 1 and three independent chains for Set 2, to check the consistency among multiple independent runs.

We used the Area Under Curve (AUC) statistic [area under the receiver-operating characteristic curve (ROC) curve] to assess the inference performance of iFad. For the retained Gibbs samples after the burn-in period, we calculate the mean of each entry of matrices Z_1 and Z_2, and the ROC curve and AUC values by choosing different cutoffs and comparing the results with the true matrices Z_1 and Z_2. The ‘ROCR’ package was used for this analysis.

For general factor analysis models, there is a scale identifiability problem associated with the loading matrix W and factor matrix X, as after integrating out the latent factors, the complete density of the observed data matrix Y is a normal distribution with covariance matrix $W\Sigma_x W' + \Psi$ (Pournara and Wernisch, 2007). In order to avoid this issue, for data simulation, we set matrix Σ_x to the identity matrix. We then tested eight combinations of Gamma parameters (α_g, β_g, α_w, β_w; Table 1) on two simulated datasets (as described in Table 2). We then asked how the result would be affected if η_0 and η_1 used for inference are different from that used in simulation. Hence, we chose $\eta_0 = \eta_1 = 0.2$ for matrix L_1, $\eta_0 = 0.15$, $\eta_1 = 0.05$ for matrix L_2 and tested the inference algorithm again on dataset 2,

Table 2. Data simulation for choosing model parameters

Set	K	G_1	G_2	J	η_0 and η_1		Density			
					π_1	π_2	L_1	L_2	Z_1	Z_2
1	18	50	50	20	0.2, 0.2	0.3, 0.1	0.316	0.167	0.394	0.38
2	15	100	50	20	0.2, 0.2	0.35, 0.05	0.05	0.0067	0.235	0.353

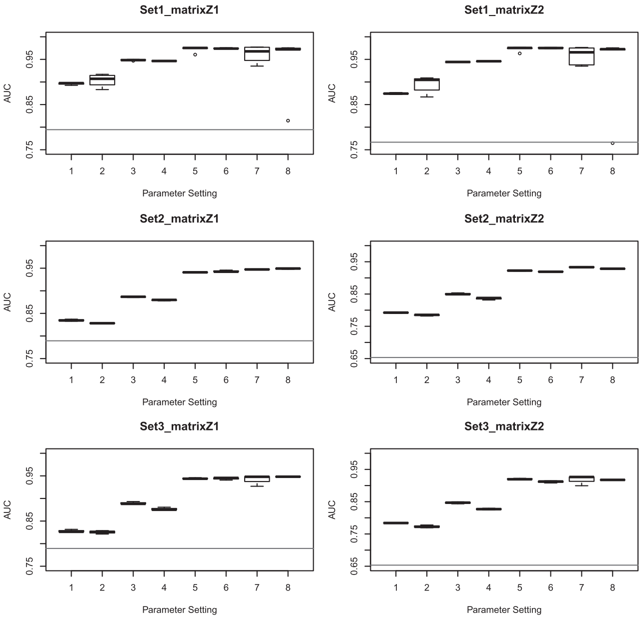


Fig. 2. AUC result of eight parameter settings for two simulated datasets. Red line is the percentage of original matching between matrices L and Z

which is denoted as ‘Set 3’. The results are shown in Figure 2. The red lines are the proportion of entries in matrix Z that have the same value with matrix L, representing the accuracy of prior knowledge.

Comparing the AUC of eight parameter settings, it can be seen that $\sigma_w = 1$ usually gives more consistent result among independent chains than putting a Gamma prior on τ_w. Nevertheless, it is obvious that α_g and β_g are the major factors here. Parameter setting 6 ($\alpha_g = 1$, $\beta_g = 0.01$ and $\sigma_w = 1$) gives best result in this comparison and is used for the remaining analysis in this article. For matrix L_2 in Set 3, $\eta_0 = 0.35$ during simulation but 0.15 during the Gibbs sampling. We checked the overlap of the inferred non-zero entries of matrix L_2 between Sets 2 and 3 (by using cutoff 0.5 to dichotomize the posterior mean for each entry). For all the eight parameter settings, the non-zero entries inferred using $\eta_0 = 0.15$ are almost always a subset of those inferred using $\eta_0 = 0.35$ (as shown in Supplementary Fig. S1).

3.1.2 Data simulation with various patterns for model performance evaluation After determining the appropriate model parameters, we explored how different dataset properties (e.g. sample size, confidence in the prior link matrix L, density of matrix L/Z and noise level) may influence the performance and robustness of the iFad model. Therefore, we simulated five other groups of datasets to investigate the effects of η_0 and η_1, density of the connectivity matrix, imbalanced dimension between the two feature spaces ($G_1 \neq G_2$), signal-to-noise ratio (SNR) and sample size (Table 3). Regarding the simulation, matrices L_1 and L_2 are randomly generated with specified density (proportion of non-zero entries). Matrices Z_1 and Z_2 are simulated based on L_1 and L_2 with Bernoulli probability specified by η_0 and η_1. The density of Z shown in Table 3 is the average value of Z_1 and Z_2. For the SNR (Group 4), the variance of each noise term (τ_g^{-1}) was calculated as $\tau_g^{-1} = \text{Var}\,(WX[g,])/\text{SNR} = \text{K/SNR}$. Three independent chains were run for each dataset in Groups 1–4, with total iteration = 30 000 and the first half as burn-in. Gibbs samples were recorded every 10th iteration. AUC results are shown in Figure 3. For Group 5, the chain usually converges slower with sample size increasing, so we tried total iteration = 10 000 (burn-in = 8000), 60 000 (burn-in = 40 000) and 100 000 (burn-in = 70 000). The AUC is plotted in Figure 4.

Table 3. Data simulations with different properties

Set	K	G_1 and G_2	J	η_0 and η_1	Density		α_g, β_g
					L	Z	
Group 1: the effect of different η_0 and η_1							
1	20	100	15	0.2	0.1	0.25	1, 0.01
2	20	100	15	0.4	0.1	0.41	1, 0.01
3	20	100	15	0.3	0.1	0.34	1, 0.01
4	Same data as Set3, but used η_0 and $\eta_1 = 0.2$ for Gibbs sampling						
5	Same data as Set3, but used η_0 and $\eta_1 = 0.4$ for Gibbs sampling						
Group 2: the effect of density of matrices L and Z							
1	20	100	15	0.2	0.01	0.21	1, 0.01
2	20	100	15	0.2	0.1	0.25	1, 0.01
3	20	100	15	0.2	0.3	0.38	1, 0.01
4	20	100	50	0.2	0.5	0.51	1, 0.01
5	20	100	50	0.2	0.7	0.61	1, 0.01
Group 3: the effect of imbalanced datasets							
1	20	100	15	0.2	0.1	0.25	1, 0.01
2	20	125, 75	15	0.2	0.1	0.26	1, 0.01
3	20	150, 50	15	0.2	0.1	0.26	1, 0.01
Group 4: the effect of SNR							
							SNR
1	20	100	15	0.2	0.1	0.257	2.5
2	20	100	15	0.2	0.1	0.260	5
3	20	100	15	0.2	0.1	0.257	10
4	20	100	15	0.2	0.1	0.271	100
5	20	100	15	0.2	0.1	0.247	500
6	20	100	15	0.2	0.1	0.247	1000
Group 5: the effect of sample size							
1	10	50	10	0.25	0.05	0.262	1, 0.01
2	10	50	30	0.25	0.05	0.251	1, 0.01
3	10	50	50	0.25	0.05	0.247	1, 0.01
4	10	50	70	0.25	0.05	0.244	1, 0.01

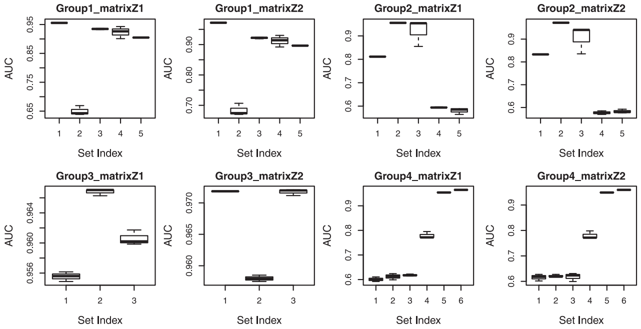


Fig. 3. AUC result of simulated data, Groups 1–4

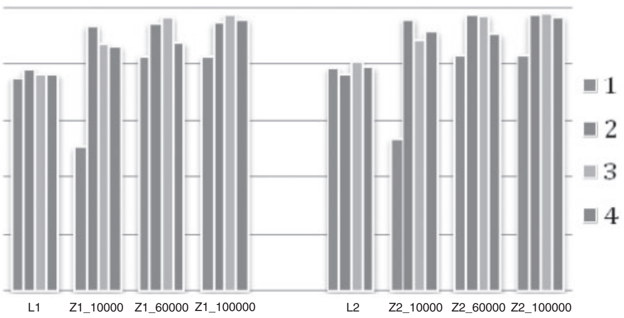


Fig. 4. AUC result of simulated data, Group 5

From these five groups of comparisons, it can be observed that

(1) When prior information about the connectivity structure matrices Z_1 and Z_2 is fairly accurate, for example, when η_0 and $\eta_1 \leq 0.3$, the AUC statistics are very good (Group 1, Sets 1 and 3), even if the η_0 and η_1 used for the Gibbs sampling algorithm deviated from the true values (Group 1, Sets 4 and 5). However, when η_0 and η_1 are large (Group 1, Set 2), the inference results are not satisfactory.

(2) iFad performs best when the density of matrices L and Z is around 0.1–0.3 (Group 2, Sets 2 and 3); too sparse (Group 2, Set 1) or too dense (Group 2, Sets 4 and 5) connectivity structure can hamper the inference result.

(3) Imbalanced dimension of the two datasets does not influence the inference result of iFad. All datasets in Group 3 achieved very good AUC statistics.

(4) The SNR has an important effect. A large SNR is desired for iFad to perform well (compare Group 4, Sets 1–3 with Sets 4–6).

(5) When sample size J is smaller than the total number of latent factors K, iFad cannot make accurate inference no matter how long the chain is run (Group 5, Set 1). Nevertheless, $J = 30$ seems to be adequate for datasets with $K = 10$ and $G_1 = G_2 = 50$ (Group 5, Set 2). Increasing sample size can improve the performance of iFad (compare Sets 3 and 4 with Set 2) but requires more iterations of the Gibbs sampling.

3.2 Application of iFad: analysis of NCI-60 datasets for drug-pathway association discovery

We then applied iFad to the joint analysis of gene expression and drug sensitivity profiles of the NCI-60 cell lines. The NCI-60 project represents a comprehensive resource for various types of ‘Omics’ characterization of 60 human cancer cell lines with nine different tissue types, including RNA expression, DNA fingerprinting, DNA methylation, sequence mutation, as well as treatment response to >100 000 compounds. The gene expression and drug sensitivity data were downloaded from the CellMiner database (Shankavaram *et al.*, 2009), with URL http://discover.nci.nih.gov/cellminer. We used ‘RNA: Affy HG-U133 (A,B)’ (44 000 probeset 2-chip set, Guanine Cytosine Robust Multi-Array Analysis (GCRMA) normalization) and ‘Drug: A4463’ for analysis.

3.2.1 Gene data preprocessing We only used the HG-U133A chip and converted probe expression to gene expression by taking the average of the probes mapped to the same gene, resulting in a total of 12 980 genes measured across 59 cell lines (expression data of the cell line ‘LC:NCI_H23’ was unavailable). The expression data were then standardized so that for each gene, mean = 0 and SD = 1 across the 59 cell lines. As we are mainly interested in the analysis of drug response-related genes, we only kept genes that are included in either of the following two lists: first, 766 cancer-related genes (Chen *et al.*, 2008); second, 8919 genes from the Integrated Druggable Genome Database Project (Hopkins and Groom, 2002; Russ and Lampel, 2005), downloaded from http://www.sophicalliance.com/. After this filtering, 6958 genes were retained.

3.2.2 Drug data preprocessing The drug data are the -log10(GI$_{50}$) values of Sulforhodamine assay for 4463 molecules (also known as the standard agents) that have known 2D structure and have been tested at-least two times. Higher values equate to higher sensitivity of cell lines. The data are also scaled so that mean = 0 and SD = 1 for each drug. Among these 4463 molecules, we only kept the 101 drugs annotated in the CancerResource database (Ahmed *et al.*, 2011). Little information is available about the targets or mechanisms of action for the other drugs.

3.2.3 Pathway association information Gene-pathway and drug-pathway association data were retrieved from the KEGG MEDICUS database (Kanehisa *et al.*, 2010). The link is http://www.genome.jp/kegg/catalog/pathway_dd.html. We compiled a list of 58 pathways that are either known to be related to cancer or have drug targets. Among the 6958 genes selected in Section 3.2.1, 1863 genes are covered by these 58 pathways and constitute the final list of genes in our real data analysis. Therefore, matrix L_1 is a binary one with dimension 1863 × 58, and the dimension of matrix L_2 is 101 × 58.

Our research objective here is to infer unknown drug-pathway associations to help better understand the mechanism of action of less well-studied compounds. We treated the pathway activity levels as latent factors in the iFad model, the gene expression data as matrix Y_1 and drug sensitivity as matrix Y_2. We compiled a list of 58 pathways (Supplementary Table S1), 1863 genes and 101 drugs for analysis, as described earlier. Gene expression data are available for 59 cell lines, representing nine different cancer types (Supplementary Table S2). Since there are only two cell lines from prostate cancer, we excluded this panel, with eight cancer types remaining for study. iFad was applied to each type, respectively,

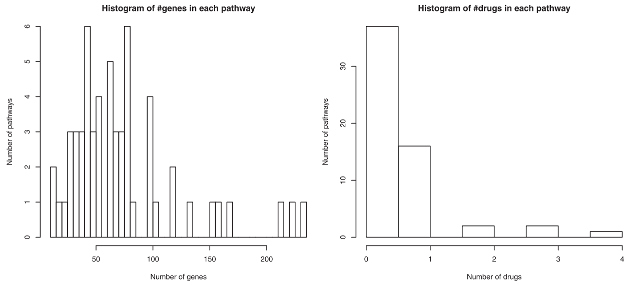


Fig. 5. Histogram of the number of genes or drugs known to be associated with the 58 KEGG pathways as *a priori*

Table 4. Number of newly inferred non-zero entries in matrix Z_2 using $\eta_0 = 0.05$ and various posterior probability cutoff values

Cutoff	0.2	0.25	0.3	0.35	0.4	0.45	0.5
BR	248	181	129	99	73	48	39
CNS	203	119	86	59	39	26	21
CO	276	209	167	148	121	98	81
LC	321	238	192	157	123	102	94
LE	200	122	86	66	53	41	33
ME	339	280	204	169	139	118	99
OV	261	191	154	118	96	78	63
RE	287	208	172	142	119	101	81
Union	1685	1282	1016	838	684	563	476

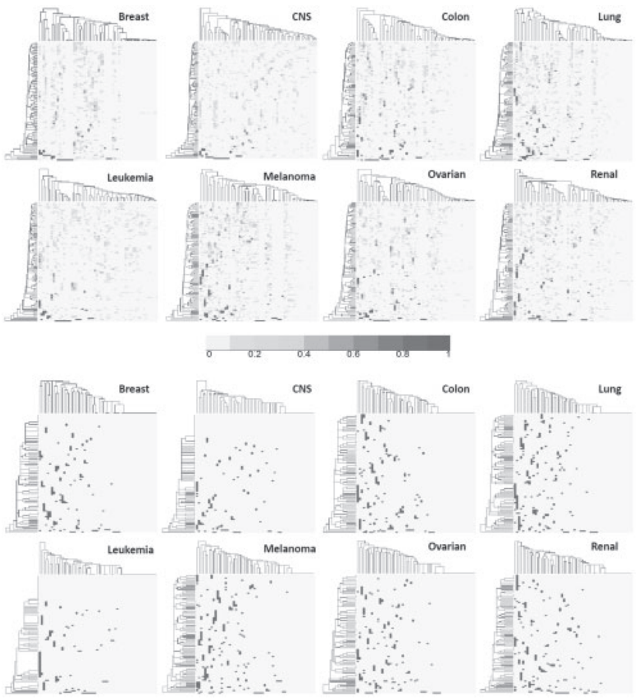


Fig. 6. Heatmaps showing the inferred drug-pathway association patterns. The upper panel shows the posterior mean for each entry of matrix Z_2. The lower panel is the dichotomized value using cutoff = 0.3. Rows correspond to drugs and columns correspond to pathways

instead of using all 57 cell lines altogether, in order to avoid potential problems arising from severe cell type heterogeneity (we checked for several well-known gene–drug correlations and found that the correlation coefficient is usually much more significant when calculated using cell lines of the same type, rather than all the 57 cell lines).

For the NCI-60 analysis, the total iteration was set to 100 000 and burn-in = 70 000. For model parameters, we set $\eta_0 = \eta_1 = 0$ for matrix π_1, because of high confidence in the gene-pathway association information from KEGG. For matrix π_2, we set $\eta_1 = 0$ and tried $\eta_0 = 0.05$, 0.1, 0.15, 0.2, 0.25, in order to infer unknown drug-pathway associations for various densities of matrix Z_2. Based on prior knowledge from the KEGG database, matrix L_1 has a density of 3.95%, whereas L_2 has a density of 0.51%. Figure 5 shows the distribution of the number of genes/drugs associated with each pathway for matrix L_1/L_2.

We compared the distribution of the posterior means of the entries of matrix Z_2, inferred using different values of η_0, as shown in the histograms of Supplementary Figure S2 and the quantile plots in Supplementary Figure S3. As expected, with η_0 increasing, the distribution of the posterior mean of Z_2 shifts to the right. For $\eta_0 =$ 0.05, the number of newly inferred non-zero entries in matrix Z_2 is shown in Table 4 based on various cutoffs.

We took a further look at the results obtained from cutoff = 0.3, because the posterior means of non-zero entries of matrix Z_2 can reach around 0.5 (for more details, see Supplementary Table S3) at this cutoff. Figure 6 shows the association pattern in the heatmap. It can be seen that the drug-pathway interaction pattern exhibits strong cell type specificity, demonstrating the importance of conducting the analysis by separating cell line groups rather than on the entire NCI-60 panel. Supplementary Figure S4 shows the total number of drug-pathway associations in barplots. Generally speaking, the cell lines of ‘non-small cell lung cancer’ and ‘melanoma’ discovered more novel drug-pathway associations, whereas ‘leukemia’, central nervous system (‘CNS’) and ‘breast cancer’ cell lines inferred fewer new associations. One possible explanation is the difference in sample size: there are nine cell lines for ‘non-small cell lung cancer’ and ‘melanoma’, but only six cell lines for ‘leukemia’, ‘CNS’ and ‘breast cancer’.

We then checked whether the newly inferred drug-pathway associations are supported by biological knowledge, based on the CancerResource database (Ahmed *et al.*, 2011) and PubMed. We chose the CancerResource database as a reference because of its comprehensiveness: it integrates drug-target information from several well-known databases, including CTD (Davis *et al.*, 2009), PharmGKB (Hernandez-Boussard *et al.*, 2008), TTD (Zhu *et al.*, 2011), DrugBank (Wishart *et al.*, 2008), as well as its own literature mining. It is worth noting that the current catalog of drug-target information is still far from complete, and the absence of specific drug-pathway associations in one database does not exclude the possibility that the interaction actually exists. Herein, we checked the inferred drug-pathway interactions using cutoff = 0.9 (with the complete list provided in Supplementary Table S4) and found that many of these associations can be validated by the CancerResource database (Table 5). For the unconfirmed associations, we checked the database CTD and found several additional validations. For example, among the ‘colon cancer’ cell line panel, drug ‘daunorubicin’ is associated with pathway

Table 5. Drug-pathway associations inferred using iFad which have been confirmed by the CancerResource database (cutoff = 0.9)

KEGG pathway	Drug	Posterior probability	Cell line panel
Glutathione metabolism	Vincristine	0.9987	LC
ErbB signaling pathway	Mitoxantrone	0.9937	LC
Thyroid cancer	Doxorubicin	0.991	RE
Glutathione metabolism	6-Mercaptopurine	0.9867	RE
Bladder cancer	Tamoxifen	0.982	LC
VEGF signaling pathway	Carmustine	0.9803	RE
Thyroid cancer	Doxorubicin	0.9713	OV
ErbB signaling pathway	Camptothecin	0.9583	LC
Bladder cancer	Edelfosine	0.958	LC
Bladder cancer	Chlorambucil	0.9473	RE
Melanoma	Chlorambucil	0.947	ME
VEGF signaling pathway	6-Mercaptopurine	0.932	ME
Bladder cancer	Geldanamycin	0.9273	CO
Thyroid cancer	Dactinomycin	0.9263	OV
Apoptosis	Thymidine	0.923	CO
Cell cycle	Tiazofurin	0.919	LC
Drug metabolism—other enzymes	Daunorubicin	0.9157	LC
VEGF signaling pathway	Lomustine	0.9103	RE
Focal adhesion	Geldanamycin	0.91	BR
Endometrial cancer	Doxorubicin	0.909	CO
VEGF signaling pathway	Quinacrine	0.9023	RE
Base excision repair	Decitabine	0.9017	LC

BR, breast; CNS, central nervous system; CO, colon; LC, non-small cell lung cancer; LE, leukemia; ME, melanoma; OV, ovarian; RE, renal.

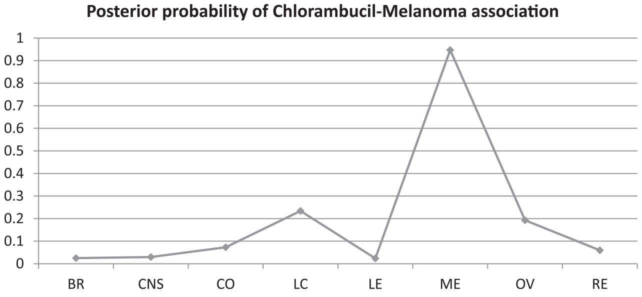


Fig. 7. Cell line specificity of the ‘chlorambucil’—‘melanoma pathway’ association

‘endometrial cancer’ with a posterior probability of 0.9547. This association is not documented in the CancerResource database, but can be confirmed by CTD. Although some drug-pathway associations are significant in more than one cell line panels (e.g. ‘doxorubicin’ acts on the ‘thyroid cancer pathway’ in both renal and ovarian cancer cell lines), most associations are still context-specific, for instance, ‘chlorambucil’ is associated with ‘melanoma pathway’ mainly in melanoma cell lines, with a high posterior probability of 0.947. In contrast, this probability is much smaller in the other cell line panels (Figure 7).

We further investigated the loading matrices W_1 and W_2 for the ‘melanoma pathway’. Since there may be sign-flip during the Gibbs sampling iterations, we calculated the posterior mean of the absolute value for each entry of matrices W_1 and W_2 after the burn-in period. Figure 8 shows the heatmap of the estimated loadings of factor ‘melanoma pathway’ on its associated 63 genes (the left part) and

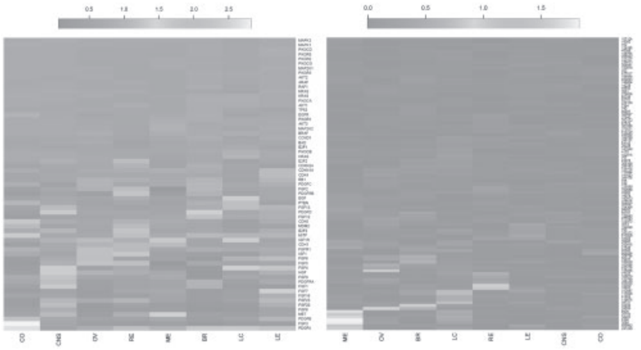


Fig. 8. Posterior mean of the absolute value of matrices W_1 and W_2 (only showing the result corresponding to the ‘melanoma pathway’) for the NCI-60 data analysis, plotted by each cell line panel

on the 101 drugs (the right part). It can be clearly observed that when the analysis is applied to the ME panel, the factor ‘melanoma pathway’ has significant loadings on several drugs; however, these drug-pathway associations are much less evident when the analysis was performed on the other cell line types.

4 DISCUSSION

Drug-target identification is one important problem in translational bioinformatics, as well as a crucial step in the early stage of drug discovery and development. Although there exist many different high-throughput technologies for molecular phenotype profiling, how to perform knowledge-based, informative data integration remains a major challenge. In this article, we have proposed a Bayesian sparse factor analysis model, iFad, for the joint analysis of gene expression and drug sensitivity profiles measured on the same set of cell lines. The aim is to identify the target biological pathways for drugs with unclear mechanism of action. This model allows natural incorporation of prior knowledge about the connectivity structure of biological pathways (e.g. KEGG pathway), and simultaneously relates the underlying pathway activity to both gene expression levels and drug response. Due to this sparsity formulation, the sample size needed to achieve satisfactory inference result can be much smaller than the number of features in either dataset. We demonstrate the performance of iFad first using simulation and then on the NCI-60 datasets. Real data analysis shows that our method is able to identify many cancer type-specific drug-pathway associations. One direction of great interest for future study is how to speed up the computation process, since MCMC methods are usually time-consuming when applied to high-dimensional inference.

Joint modeling of expression profiles and drug-related data represent an increasingly important and popular trend in the future. Besides the bi-clustering method ISA mentioned in Section 1 (Kutalik *et al*., 2008), another seminal work in this field (Chang *et al*., 2005) used Bayesian networks to model the gene–drug dependency, also on the NCI-60 data. Due to computational constraints of Bayesian network models, extensive feature selection was performed before the network inference. A more recent work (Chen *et al*., 2009) developed a linear regression model that integrates genotype and gene expression data generated under drug-free conditions of yeast segregants to predict the response to various drugs. From a statistical point of view, joint analysis of paired datasets can be achieved using a number of techniques, such as canonical correlation, bipartite graph inference, model-based clustering, etc. With the availability of more and more types of high-throughput datasets from the same panel of samples, novel statistical methods are in great need for knowledge-guided combined analysis.

Funding: National Institutes of Health (GM59507 to H.Z.) and NIH R21-GM084008 to Ning Sun.

Conflict of Interest: None declared.

REFERENCES

Ahmed,J. *et al.* (2011) CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res.*, **39**, D960–D967.

Bader,G.D. *et al*. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504–D506.

Boyce,S.E. *et al.* (2009) Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.*, **394**, 747–763.

Bussey,K.J. *et al.* (2006) Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol. Cancer Ther.*, **5**, 853–867.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Chang,J.H. *et al.* (2005) Bayesian network learning with feature abstraction for gene-drug dependency analysis. *J. Bioinform. Comput. Biol.*, **3**, 61–77.

Chen,B.J. *et al.* (2009) Harnessing gene expression to identify the genetic basis of drug resistance. *Mol. Syst. Biol.*, **5**, 310.

Chen,J. *et al.* (2008) Genomic profiling of 766 cancer-related genes in archived esophageal normal and carcinoma tissues. *Int. J. cancer*, **122**, 2249–2254.

Czodrowski,P. *et al.* (2009) Computational approaches to predict drug metabolism. *Expert Opin. Drug Metab. Toxicol.*, **5**, 15–27.

Davis,A.P. *et al.* (2009) Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.*, **37**, D786–D792.

Ecker,G.F. *et al*. (2008) Computational models for prediction of interactions with ABC-transporters. *Drug Discov. Today*, **13**, 311–317.

Gharib,S.A. *et al.* (2006) Computational identification of key biological modules and transcription factors in acute lung injury. *Am. J. Respir. Crit. Care Med.*, **173**, 653–658.

Hernandez-Boussard,T. *et al.* (2008) The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.*, **36**, D913–D918.

Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.

Ikediobi,O.N. *et al.* (2006) Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Mol. Cancer Ther.*, **5**, 2606–2612.

Irwin,J.J. *et al.* (2009) Automated docking screens: a feasibility study. *J. Med. Chem.*, **52**, 5712–5720.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Kolb,P. *et al.* (2009) Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.*, **20**, 429–436.

Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.

Kutalik,Z. *et al.* (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.*, **26**, 531–539.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Meng,J. *et al.* (2011) Bayesian non-negative factor analysis for reconstructing transcription factor mediated regulatory networks. *Proteome Sci.*, **9** (Suppl 1), S9.

Nigsch,F. *et al.* (2009) Computational toxicology: an overview of the sources of data and of modelling methods. *Expert Opin. Drug Metab. Toxicol.*, **5**, 1–14.

Pournara,I. and Wernisch,L. (2007) Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, **8**, 61.

Pujol,A. *et al.* (2010) Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.*, **31**, 115–123.

Russ,A.P. and Lampel,S. (2005) The druggable genome: an update. *Drug Discov. Today*, **10**, 1607–1610.

Schadt,E.E. *et al*. (2009) A network view of disease and compound screening. *Nat. Rev. Drug Discov.*, **8**, 286–295.

Shankavaram,U.T. *et al.* (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol. Cancer Ther.*, **6**, 820–832.

Shankavaram,U.T. *et al.* (2009) CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics*, **10**, 277.

Sharma,S.V. *et al.* (2010) Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat. Rev. Cancer*, **10**, 241–253.

Sharp,K. *et al.* (2010) A comparison of inference in sparse factor analysis. *Submitted to the J. Mach. Learn. Res.*, September.

Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.

Staunton,J.E. *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA*, **98**, 10787–10792.

West,M. (2003) Bayesian factor regression models in the "large *p*, small *n*" paradigm. In Bernardo,J.M. *et al.* (eds) *Bayesian Statistics*, Oxford University Press, Oxford, UK, Vol. **7**, pp. 733–742.

Wishart,D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.

Yeh,H.Y. *et al.* (2009) Identifying significant genetic regulatory networks in the prostate cancer from microarray data based on transcription factor analysis and conditional independency. *BMC Med Genomics*, **2**, 70.

Yeh,P. *et al.* (2006) Functional classification of drugs by properties of their pairwise interactions. *Nat. Genet.*, **38**, 489–494.

Yildirim,M.A. *et al.* (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.

Yu,T. and Li,K.C. (2005) Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics*, **21**, 4033–4038.

Zavodszky,M.I. and Kuhn,L.A. (2005) Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci.*, **14**, 1104–1114.

Zhu,F. *et al.* (2011) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.*, **40**, D1128–D1136.