BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 20 2012, pages 2662–2670
doi:10.1093/bioinformatics/bts502

Data and text mining

Advance Access publication August 24, 2012

FacPad: Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment

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

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

Associate Editor: Martin Bishop

ABSTRACT

Motivation: It is well recognized that the effects of drugs are far beyond targeting individual proteins, but rather influencing the complex interactions among many relevant biological pathways. Genome-wide expression profiling before and after drug treatment has become a powerful approach for capturing a global snapshot of cellular response to drugs, as well as to understand drugs' mechanism of action. Therefore, it is of great interest to analyze this type of transcriptomic profiling data for the identification of pathways responsive to different drugs. However, few computational tools exist for this task.

Results: We have developed FacPad, a Bayesian sparse factor model, for the inference of pathways responsive to drug treatments. This model represents biological pathways as latent factors and aims to describe the variation among drug-induced gene expression alternations in terms of a much smaller number of latent factors. We applied this model to the Connectivity Map data set (build 02) and demonstrated that FacPad is able to identify many drug–pathway associations, some of which have been validated in the literature. Although this method was originally designed for the analysis of drug-induced transcriptional alternation data, it can be naturally applied to many other settings beyond polypharmacology.

Availability and implementation: The R package 'FacPad' is publically available at: http://cran.open-source-solution.org/web/packages/FacPad/

Contact: hongyu.zhao@yale.edu

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

Received on May 7, 2012; revised on August 1, 2012; accepted on August 7, 2012

1 INTRODUCTION

One of the common features of complex diseases (such as cancer, cardiovascular diseases and neurological disorders) is the involvement of many genes and biological pathways during the pathogenic process. On the other hand, human physiological systems show different degrees of robustness against single-point perturbation due to functional redundancy, various feedback mechanisms and other immune response. Therefore, successful treatment of complex diseases requires 'polypharmacology', which aims to design multi-targeting therapeutics and may represent a new paradigm shift in drug discovery (Xie *et al.*, 2012).

Comprehensive and accurate understanding of the targeting spectrum of various drugs is the prerequisite for practical application of polypharmacology. Although the molecular mechanisms of some drugs are well studied, unexpected side effects or toxicity of a large number of preclinical and clinical drugs are often observed (Campillos *et al.*, 2008; Xie *et al.*, 2012). Therefore, the construction of genome-wide drug–target network is a problem of great interest in the bioinformatics community and beyond.

Current approaches for computational drug target identification can be roughly categorized into three groups: ligand, phenotype or target based (Xie *et al.*, 2012). Ligand-based approaches assume that similarity in chemical structures between drugs indicates similar targeting activities. A well-known example in this category is the QSAR method (Quantitative Structure Activity Relationship), which uses 2D topological feature vectors (which usually encode atom types and their bonding structure) of drugs to train machine learning models in order to predict their binding activity toward specific target proteins (Xie *et al.*, 2012; Yamanishi *et al.*, 2010). However, decent performance of ligand-based approach requires a large enough number of known ligands for target proteins of interest, which may be difficult to meet in practice.

A second class of methods associates different drugs by comparing the biological phenotype responses, such as cell lines' gene expression profiles or proteomic data (Xie *et al.*, 2012). Seminal work in this direction includes the national NCI-60 project (Shoemaker, 2006), which screened 60 human tumor cell lines against more than 100 000 compounds and constructed a public repository for the basal gene expression and drug sensitivity information. The Connectivity Map project initiated by the Broad Institute went a step further (Lamb, 2007; Lamb *et al.*, 2006). Although it only focused on five human cancer cell lines, the project generated genome-wide expression profiles both before and after drug treatment for 1309 compounds. In this way, compounds can be connected into a network by comparing their ranked lists of up- and down-regulated genes (Iorio *et al.*, 2009, 2010). Other phenotype information such as cell imaging and side effects have also been utilized to associate different drugs and to make inference about their potential targets (Campillos *et al.*, 2008; Young *et al.*, 2008).

The third category of methods looks at the similarity between target proteins to predict drug–target interactions, based on structure, sequence, evolutionary as well as functional information (Xie *et al.*, 2012). Several recent studies integrate sequence features of targets with ligand fingerprints to train statistical

*To whom correspondence should be addressed.

learning models for drug–target interaction prediction (He *et al.*, 2010; Nagamine *et al.*, 2009; Vina *et al.*, 2009; Yamanishi *et al.*, 2008). Another well-known method in this category is docking analysis, which predicts the preferred orientation of drug candidates to potential target proteins when they are bound to each other to form a stable complex (Kitchen *et al.*, 2004). However, docking cannot be applied to proteins whose 3D structures are unknown. Therefore, it is difficult to use the docking on a genome-wide scale (Yamanishi *et al.*, 2010).

The method proposed in this article falls into the second category, phenotype-based approaches. It is designed for the analysis of genome-wide transcriptional response profiles upon screening of a large number of compounds, which shares similar data structure with the Connectivity Map project. Previous studies on treatment response data usually focus on differential expression analysis: to identify the most up- and down-regulated genes for each drug and then check for pathway enrichment of these response signatures (Iskar *et al.*, 2010; Smalley *et al.*, 2010) or to define certain drug similarity metrics based on the ranked gene list as the Connectivity Map project (Iorio *et al.*, 2010; Lamb *et al.*, 2006). Drawbacks with the former approach include that it only utilizes the information of the most differentially expressed (DE) genes while disregarding all the other data points. Moreover, the choice of the DE cutoff is subjective, leading to potentially inconsistent pathway enrichment results. The Connectivity Map approach mainly aims at assessing the similarity between different drugs or the construction of a 'drug-network', rather than direct identification of target biological pathways of various compounds.

Target pathway identification is a critical step for therapeutics design in the age of systems pharmacology (Zhao and Iyengar, 2012). Compared with traditional methods focusing on the inference of separate target genes, taking the pathway perspective allows a more integrative understanding of the mechanism of action and physiological effects of compounds. In order to address the aforementioned limitations of existing analysis procedures, we propose 'FacPad', which applies a Bayesian sparse factor model to explain the gene-wise treatment response in terms of latent biological pathway activity changes upon individual treatment. By encoding pathways as latent factors, FacPad naturally incorporates prior knowledge on pathway–gene association structure to aid the inference on drug targets.

The idea of FacPad is conceptually related to one of our previous studies 'iFad', which stands for an integrative factor analysis model for drug–pathway association inference (Ma and Zhao, 2012). However, distinct from FacPad, iFad aims to jointly analyze two phenotype profiles measured across the same set of samples, that is, two matrices with the same number of columns (sample size) but different number of rows (dimension of the feature space). The two data matrices are decomposed using a common set of latent factors, which represent biological pathways. Therefore, iFad can be applied to the analysis of paired transcription profiles (basal level gene expression, before drug treatment) and drug sensitivity data ('GI_{50}' values, drug concentrations required to inhibit growth by 50%) across the same cell lines, in order to infer the gene–pathway and drug–pathway associations simultaneously.

Due to its general factor model framework, FacPad can be easily applied to many other settings apart from drug-induced pathway response data. From the perspective of data mining, this method is basically performing dimension reduction analysis: it aims to explain the variation in the original observed variables in terms of a much smaller number of latent factors. Different from traditional factor analysis or PCA, FacPad is more convenient to use when there exist certain prior knowledge on the association pattern between the latent factors and the observed independent features.

The remaining of this article is structured as follows: we will first describe the statistical model of FacPad in Section 2; in Section 3, application of the model on the Connectivity Map data set is presented; the article concludes with a brief discussion on issues worth of note and potential future work.

2 METHODS

2.1 Model description

FacPad is designed for the analysis of a $G \times J$ data matrix Y, which describes the genome-wide transcriptional response upon different treatments. G is the total number of genes measured by the microarray platform (or proteins if proteomics data is utilized), whereas J is the number of treatments which have been screened. Each treatment is usually a specific drug at a given dosage with a specific treatment time. Therefore, the number of unique drugs tested is often smaller than the number of treatments. Each entry in matrix Y is the response value of a single gene upon a certain treatment and is usually computed as the ratio (or fold change) of the gene expression levels after versus before treatment. Prior information on pathway structure is represented as a $G \times K$ binary matrix L, where K is the number of pathways included in the analysis. This information can be retrieved from many public pathway databases such as KEGG (Kanehisa *et al.*, 2012; Kotera *et al.*, 2012), MetaCyc (Caspi *et al.*, 2012) and so on (Bader *et al.*, 2006). If the gth gene belongs to the kth pathway, $L[g, k] = 1$ and it takes value 0 otherwise.

In order to make inference on the target pathways of different treatments, FacPad applies a Bayesian sparse factor model for the decomposition of matrix Y, which was originally proposed for microarray data analysis (West, 2003). This type of model is well suited for the research objective here: first, under the Bayesian setting, prior knowledge on pathway structures can be easily utilized to guide the inference process; second, a sparse model represents a more realistic description of the association pattern between pathways and genes. The outline of the model is as follows:

$$
\begin{aligned}
& Y = WX + E, E_{\bullet,j} \sim N(0, \Sigma) \\
& \Sigma = diag\{\tau_1^{-1}, \tau_2^{-1}, \ldots, \tau_g^{-1}, \ldots, \tau_G^{-1}\} \\
& X_{k,j} \sim Normal(0, 1) \\
& \tau_g \sim Gamma(\alpha_g, \beta_g), g = 1, 2, \ldots G \\
& W_{g,k} = \begin{cases} 0, & if \quad L_{g,k} = 0 \\ N(0, \tau_w^{-1}), & if \quad L_{g,k} = 1 \end{cases} \\
& \tau_w \sim Gamma(\alpha_w, \beta_w)
\end{aligned}
$$

Matrix W is the $G \times K$ factor loading matrix describing the association strength between the genes and the pathways. Most entries in matrix W are 0, with the sparsity structure defined exactly the same as matrix L. For each non-zero entry in matrix W, $W_{\mathrm{g,k}}$, we put a normal prior with mean 0 and precision τ_w. τ_w can be either set to a constant or assumed to follow a conjugate Gamma distribution. In the following analysis, we set $\tau_w = 1$. Matrix X is the $K \times J$ latent factor matrix, with each factor representing the treatment response of a specific biological pathway. Entries in matrix X are assumed to follow the standard normal distribution. Matrix E is the $G \times J$ noise matrix. Each column of E is generated from a multivariate normal distribution with mean 0 and a $G \times G$

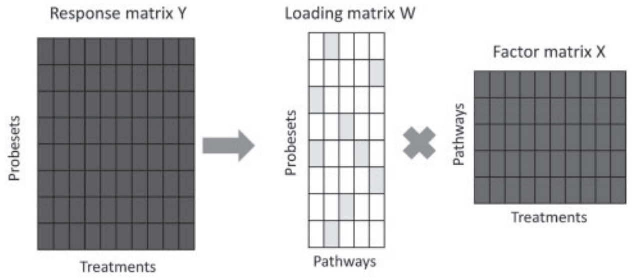
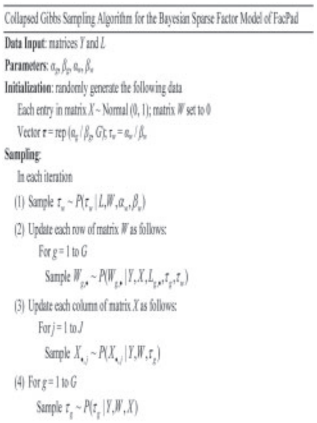


Fig. 1. Illustration of the Bayesian sparse factor model used in FacPad



```
Collapsed Gibbs Sampling Algorithm for the Bayesian Sparse Factor Model of FacPad
Data Input: matrices $Y$ and $L$
Parameters: $\alpha_g, \beta_g, \alpha_w, \beta_w$
Initialization: randomly generate the following data
  Each entry in matrix $X$ ~ Normal (0, 1); matrix $W$ set to 0
  Vector $\tau$ = rep ($\alpha_g / \beta_g$, $G$); $\tau_w = \alpha_w / \beta_w$
Sampling:
  In each iteration
  (1) Sample $\tau_w \sim P(\tau_w | L, W, \alpha_w, \beta_w)$
  (2) Update each row of matrix $W$ as follows:
      For $g$ = 1 to $G$
        Sample $W_{g,\bullet} \sim P(W_{g,\bullet} | Y, X, L_{g,\bullet}, \tau_g, \tau_w)$
  (3) Update each column of matrix $X$ as follows:
      For $j$ = 1 to $J$
        Sample $X_{\bullet,j} \sim P(X_{\bullet,j} | Y, W, \tau_g)$
  (4) For $g$ = 1 to $G$
      Sample $\tau_g \sim P(\tau_g | Y, W, X)$
```

Fig. 2. Outline of the collapsed Gibbs sampling algorithm of FacPad

diagonal covariance matrix Σ. The precision of the gth gene, τ_g, is assumed to follow a Gamma prior with shape parameter α_g and rate parameters β_g. Under this setting, we can derive the prior probability of different components of the model, as well as the complete joint posterior probability (see the Supplementary Material for details). An intuitive illustration of the model is shown in Figure 1.

2.2 Inference algorithm

For Bayesian inference, Gibbs sampling is a widely used technique to approximate the joint distribution of models with a large number of parameters. However, standard Gibbs sampler may suffer from poor mixing due to the dependence between matrices W and Z in this case. Therefore, we have utilized a modified collapsed Gibbs sampling algorithm based on several previous studies (Pournara and Wernisch, 2007; Rattray *et al.*, 2009; Sabatti and James, 2006). The outline of the algorithm is listed below (Fig. 2). Detailed derivations of the posterior conditional distributions are provided in the Supplementary Material. We have implemented the above algorithm as the R package 'FacPad', which is publicly available on CRAN (http://cran.open-source-solution.org/web/packages/FacPad/) and also our group's website (http://bioinformatics.med.yale.edu/group/).

2.3 Data simulation for model parameter selection

In order to choose the Gamma parameters (α_g and β_g) used in the inference algorithm for the FacPad model, we simulated the binary link matrix **L**, loading matrix **W**, latent factor activity matrix **X** and the error matrix **E**, using the setting shown in Table 1. We then compared the following pairs of α_g and β_g values: (0.7, 0.3), (1, 0.1) and (1, 0.01). For each pair of Gamma parameters, we run the Gibbs sampling for 20 000 iterations and the first 5000 iterations were discarded as burn-in period. For each of the remaining 15 000 iterations, we computed the estimated matrix $\hat{Y} = \hat{W}\hat{X}$ and then took the average for each entry across the 15 000 iterations to get the final estimate for matrix **Y**. We then calculate the root-mean-squared error (RMSE) $= \sqrt{\sum (Y-\hat{Y})^2 / length(Y)}$.

Table 1. Simulated data sets for Gamma parameter selection

Matrix	Related model parameters
L	G (nrow) = 100, K (ncol) = 20, density (% of 1-entries) = 0.1
W	G (nrow) = 100, K (ncol) = 20, $\tau_w = 1$
X	K (nrow) = 20, J (ncol) = 30
E	G (nrow) = 100, J (ncol) = 30, $\sigma_g = 0.2$

2.4 Preprocessing of the connectivity map data

We applied 'FacPad' to the analysis of microarray data sets generated by the Connectivity Map project (Lamb, 2007; Lamb *et al.*, 2006). The current version of the web interface (CMap Build 02, http://www.broadinstitute.org/cmap/) provides public download of genome-wide transcriptional profiles of five human cancer cell lines (MCF7: human breast cancer; HL60: human promyelocytic leukemia; ssMCF7: MCF7 grown in a different vehicle; PC3: human epithelial prostate cancer; SKMEL5: human skin melanoma) both before and after the treatments of 1309 distinct bioactive small molecules. In all, there are 7056 Affymetrix microarrays, among which 6100 arrays are from drug-treated cell lines and the others are control samples. Three microarray platforms were used: 807 HG-U133A arrays, 220 HT_HG-U133A_EA arrays and 6029 HT_HG-U133A arrays, belonging to 302 different batches.

We downloaded the raw.CEL files of all the 7056 arrays from the CMap database (http://www.broadinstitute.org/cmap/cel_file_chunks.jsp). First, basic quality check of the.CEL data was performed using the NUSE (Normalized Unscaled Standard Error) and RLE (Relative Log Expression) metrics (Brettschneider *et al.*, 2008), which are provided in the R package 'affyPLM'. For every 25 arrays from the same platform, boxplots of NUSE and RLE were drawn and arrays with bad quality were discarded. After array filtering, we performed GCRMA normalization on the.CEL files (also within each platform, respectively) using R package 'gcrma'. Next, for each retained treatment array, the corresponding control sample was identified (if there are more than one control array, then for each probeset, the average value across all controls was computed). After removing the treatment arrays with no control, transcriptional response value for each probeset was calculated as the fold change (treatment data/control data). The final processing step was to remove potential batch effects using the function 'removeBatchEffect' of the R package 'limma'.

In the following analysis, we only used the data from the HT_HG-U133A array platform, which consists of 4466 expression response profiles, representing 1084 different compounds. More details about data from the HT_HG-U133A array platform are shown in Table 2.

The Affymetrix HT_HG-U133A array platform covers 22 277 probe sets genome-wide. Within each cell line panel, the top 5000 most varying probe sets were selected, which were then mapped to KEGG pathways using the DAVID online annotation database (Huang da *et al.*, 2009a, b). A second filtering step was applied from the pathway perspective, where only pathways with more than 10 associated probe sets were retained. Finally, probe sets not belonging to these retained pathways were excluded.

Table 2. Transcriptional response profiles from HT_HG-U133A platform

Tumor cell line	MCF7	HL60	PC3
Number of treatment response profiles	2349	746	1371
Number of different batches	100	22	75
Number of different compounds represented	1069	643	1004
Number of probe sets after filtering	1533	1606	1538
Number of KEGG pathways after filtering	123	130	126
Density of prior matrix L (% of 1-entries)	0.023	0.021	0.024

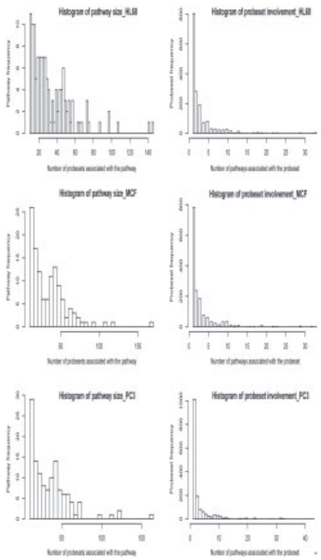


Fig. 3. Histogram describing the sparsity patterns of prior connectivity matrix L for the three different tumor cell line panels. Each row corresponds to a cell line panel. The left column shows the pathway size (number of probe sets in a certain pathway) distribution, whereas the right column shows the probe set involvement (number of pathways a certain probe set associates with) distribution

3 RESULTS

We then applied FacPad to the analysis of the treatment response profiles of three tumor cell line panels from the Connectivity Map project (Lamb, 2007; Lamb *et al.*, 2006). After preprocessing as described in Section 2, there remained 2349 treatment profiles for the MCF cell line panel, 746 profiles for the HL60 panel and 1371 profiles for the PC3 panel. For each panel, around 1500 probe sets mapped to ~130 KEGG pathways were used in the analysis (Table 2). The binary matrix L was also constructed based on these probe set—pathway associations, which in turn determines the sparsity structure of loading matrix W. Therefore, for the MCF7 cell line panel, dim (matrix Y) = 1533 × 2349, dim (matrix L) = 1533 × 123; for HL60, dim (matrix Y) = 1606 × 749, dim (matrix L) = 1606 × 130; for PC3, dim (matrix Y) = 1538 × 1371, dim (matrix L) = 1538 × 126. Figure 3 shows the sparsity pattern of matrix L of the three cell line panels.

3.1 Parameter setting for the collapsed Gibbs sampling algorithm

The Gamma parameters (α_g and β_g) were chosen based on practical experiments via simulation, as described in Section 2.3 (Table 1). The RMSE of estimated matrix **Y** for the three pairs of Gamma parameters tested are: (i) $\alpha_g = 0.7$, $\beta_g = 0.3$, RMSE = 0.09936; (ii) $\alpha_g = 1$, $\beta_g = 0.1$, RMSE = 0.09606 and (iii) $\alpha_g = 1$, $\beta_g = 0.01$, RMSE = 0.09553. Therefore, we set $\alpha_g = 1$, $\beta_g = 0.01$, for the following analysis presented in this article. For Gibbs sampling, we set the number of iterations to 30 000 for HL60 and PC3 panel (first 5000 iterations as burn-in), whereas 20 000 for MCF panel (first 7500 iterations as burn-in) since the sample size of this panel is much larger and each iteration of the Gibbs sampling takes more time. Approximately, each iteration of the Gibbs sampling takes 13 s for HL60 panel, 20 s for PC3 panel and 37 s for MCF7 panel. To reduce the effect of auto-correlation between adjacent iterations and the data storage burden, we only recorded the Gibbs sampling results every 10th iteration. The number of burn-in iterations was chosen based on the MCMC trace plot. The sample trace plots are shown in Supplementary Figure S1, which illustrate the sampled elements/entries in vector τ, matrix $\boldsymbol{X}$ and matrix $\boldsymbol{W}$ during the running of the chain. One issue with factor model is the sign flipping problem: changing the sign of matrix $\boldsymbol{X}$ and matrix $\boldsymbol{W}$ at the same time will not change the estimated value of matrix $\boldsymbol{Y}$. Since we are more interested in the inference of treatment–pathway association strength (the absolute value of matrix $\boldsymbol{X}$), rather than the sign of the correlation between probe sets and pathways or between treatments and pathways, we only plotted the absolute values of selected entries in matrix $\boldsymbol{X}$ and $\boldsymbol{W}$.

3.2 Estimated gene–pathway and treatment–pathway association relationships of MCF7, HL60 and PC3 cell line panels of connectivity map data

As mentioned earlier, in the analysis presented here, we mainly focus on the inference of the association strength between genes/treatments and the KEGG pathways, rather than the positive or negative signs of the correlation. Therefore, we estimated the absolute value of latent factor matrix $|\boldsymbol{X}|$ (treatment–pathway association) and the loading matrix $|\boldsymbol{W}|$ (gene–pathway association) by taking the average of the absolute value of each entry across the Gibbs sampling iterations (after excluding the burn-in samples). A visual representation of the inferred matrix $|\boldsymbol{X}|$ is shown in Figure 4, whereas histogram showing the distribution of inferred matrix $|\boldsymbol{X}|$ and $|\boldsymbol{W}|$ entries is shown in Supplementary Figure S2. From the heatmaps, it can be seen that for all the three tumor cell line panels, the treatment–pathway association relationships exhibit certain modular structure. It is not unusual that several treatments are strongly associated with several pathways at the same time, forming a notable bi-cluster in the heatmap.

In order to validate the inferred treatment–pathway associations, we resorted to the cataloged drug target information from the comparative toxicogenomics database (CTD; Davis *et al.*, 2011). CTD contains curated interactions between chemicals and genes/proteins. It also provides a list of pathways that are statistically enriched among the genes/proteins that interact with a chemical. The significance of enrichment was calculated by the hypergeometric distribution and adjusted for multiple testing using the Bonferroni method. We downloaded the 'chemical–pathway enriched associations' zip file from their web interface (URL: http://ctdbase.org/downloads/#chem-pathwaysenriched), updated as of March 14, 2012. The total number of drugs included in the CTD database is 143 668, among which 5460 drugs have enriched associations with at least 1 of 230 KEGG pathways. The total number of chemical–pathway interactions is 152 556, leading to a 5460 × 230 drug–pathway interaction matrix with density 12.15%.

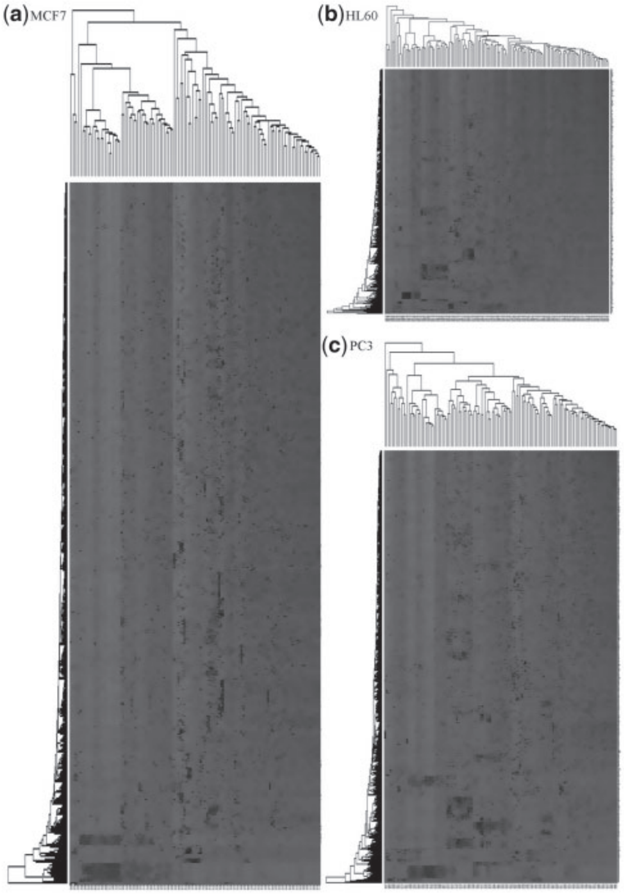


Fig. 4. Heatmaps of estimated matrix $|X|$ across the three cell line panels of the Connectivity Map data sets. Rows correspond to treatments and columns correspond to KEGG pathways. Green to red represents low to high values. **(a–c)** plots the estimated matrix $|X|$ for the cell line panels MCF, HL60 and PC3, respectively

Since the chemical information documented in CTD is incomplete, drug–pathway associations inferred by matrix $|\boldsymbol{X}|$ but not recorded in CTD may also be true interactions, rather than false positives. Therefore, the true positive rate (aka sensitivity or recall) is a more reliable performance measure here. Detailed information about this analysis is shown in Table 3. For the three estimated latent factor matrix $|\boldsymbol{X}|$, after excluding treatments not included in CTD, we checked whether larger entries in matrix $|\boldsymbol{X}|$ are more likely to indicate true (CTD-validated) drug–pathway associations, using a metric similar to the so-called 'recall enhancement' (Zhou *et al.*, 2010). First, we ranked all the entries' values of matrix $|\boldsymbol{X}|$ from largest to smallest and checked whether the corresponding drug–pathway associations are confirmed by CTD. Then, we calculated the fold enrichment of true positives compared with random selection, for increasing number of top-ranked entries. An illustration of this calculation procedure using matrix $|\boldsymbol{X}|$ derived from the HL60 panel is shown in Table 4. The last column, fold enrichment of true positives (FE_TP), is computed as follows:

$$FE\_TP = \frac{TP}{Random\_TP} = \frac{TP \times (\#All\_Entries)}{P \times (Entry\_Rank)}$$

where 'TP' stands for 'number of true positives', 'Random_TP' stands for 'expected number of true positives by random selection', P refers to the total number of positives in the whole matrix and 'Entry_Rank' stands for the number of top-ranked entries included in counting TP. We also checked the fold enrichment for false positives (FE_FP), which is defined similarly:

$$FE\_FP = \frac{FP}{Random\_FP} = \frac{FP \times (\#All\_Entries)}{N \times (Entry\_Rank)}$$

Figure 5 shows the 'FE_TP' and 'FE_FP' for each of the three cell line panels (only plotting top 10 000 entries here since the remaining part has the same ascending/descending trend). It can be seen that both the peaks of 'FE_TP' and the bottoms of 'FE_FP' are achieved simultaneously at a very top rank (top 0.0113% for HL60 panel, top 0.00136% for PC3 panel and top 0.6205% for MCF7 panel). With more and more lower-ranked entries included, 'FE_TP' gradually decreases while 'FE_FP' gradually increases, both of which finally approximate to 1 (same as random selection). Therefore, the values of matrix $|X|$ entries can fairly reflect the association strength between drugs and KEGG pathways.

Table 3. Validation analysis using CTD curated data

Cell line panel	No. of treatments included in the factor analysis	No. of treatments whose drugs are included in CTD	No. of unique drugs among these treatments	No. of KEGG pathways included in the factor analysis
MCF	2349	2082	932 (average replicates = 2.234)	123
HL60	746	681	579 (average replicate = 1.176)	130
PC3	1371	1167	869 (average replicate = 1.343)	126

Table 4. The calculation procedure for fold enrichment of true positive rate of drug–pathway interactions inferred by ranked entries in matrix |***X***|

Rank	Entry value	KEGG pathway ID	Drug	CTD confirmation (1 = true, 0 = false)	Cumulative true positives	Fold enrichment of true positives
1	20.56	4620	Suloctidil	0	0	0
2	16.31	4630	Alprostadil	0	0	0
3	15.11	830	Tretinoin	1	1	2.5096
4	15.01	5322	Anisomycin	0	1	1.8821
5	15.00	5016	Tretinoin	1	2	3.0115
...	...	...	...	...	...	...
88 528	0.09	590	Gabapentin	0	11 758	0.9999
88 529	0.09	590	Quercetin	1	11 759	1
88 530	0.09	590	Gliquidone	0	11 759	1

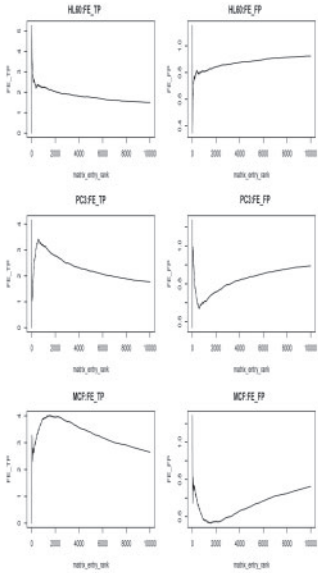


Fig. 5. Fold enrichment of true positives in matrix |***X***| with increasing number of top-ranked entries

We then checked the rank positions of different pathways for individual treatment (lower ranks indicate higher matrix |***X***| values). The results of several representative treatments for each of the three cell line panels are shown in Supplementary Tables S1, S2 and S3.

3.3 Representative treatment–pathway modules identified in the latent factor matrix *X* by FacPad

In the following, we will briefly describe a few number of representative treatment–pathway modules detected in the latent factor matrix |***X***| (notable bi-clusters in the heatmaps of Figure 4), as a demonstration of the analysis result on Connectivity Map data sets. Several selected modules are listed in Table 5. And, we will discuss one module from each of the three cell line panels below.

For the HL60 cell line panel, one module includes the pathways 'Arachidonic acid metabolism', 'Retinol metabolism' and 'Huntington's disease', with corresponding drugs 'tretinoin' and 'isotretinoin'. Both tretinoin and isotretinoin are retinoic acids, whose administration into human body will naturally result in the involvement of the above two fatty acid metabolism pathways. The 'Huntington's disease' pathway contains many genes which are potential targets of tretinoin and isotretinoin (e.g. CASP3, BAX). There have also been a number of studies demonstrating the effect of retinoids and related compounds on the treatment of neuro-inflammatory disease such as schizophrenia, Parkinson's, Huntington's and Alzheimer's diseases (Lane and Bailey, 2005; Luthi-Carter *et al.*, 2000; O'Reilly *et al.*, 2006).

For the PC3 cell line panel, one module associates the 'antigen processing and presentation pathways' with many drugs that are closely related to the immune system, such as ebselen, nifedipine, disulfiram, 15-delta prostaglandin J2, withaferin A and MG-262. The remaining drugs in this module have no or very little information on associated pathways from CTD, but may also be related to antigen processing or other immune functions based on our additional literature mining. For example, mometasone has been shown to inhibit adhesion molecular expression (an important player during immune response) in psoriasis (Berti *et al.*, 1998), as well as to have effects on mammalian allergic rhinitis model (Tsumuro *et al.*, 2005).

For the MCF7 panel, it shares the retinol pathway module with HL60 and also the antigen module with PC3. Another notable module in this panel is the one with drugs anisomycin and cephaeline. Anisomycin (a.k.a. flagecidin) is an antibiotic which can inhibit protein biosynthesis. Cephaeline is an alkaloid which has effect on stimulating the stomach lining and inducing vomiting. In this module, the two drugs are inferred to be associated with a number of signaling pathways, such as 'Toll-like receptor signaling pathway' and 'NOD-like receptor signaling pathway', which can be validated from the CTD drug–pathway interaction catalog. 'Leukocyte transendothelial migration pathway' is also relevant here since it is a necessary transitive step in most immune-related signal transduction.

3.4 Comparison with GSEA

We compared the performance of FacPad with the well-known Gene Set Enrichment Analysis, GSEA (Mootha *et al.*, 2003;

Table 5. List of representative treatment–pathway modules identified in the matrix $|X|$ for each of the three cell line panels

Pathway	Treatment
HL60 cell line panel	
Arachidonic acid metabolism Retinol metabolism Huntington's disease	Tretinoin, isotretinoin
Fc gamma R-mediated phagocytosis	Valproic acid
Complement and coagulation cascades Endocytosis Regulation of actin cytoskeleton Cell adhesion molecules (CAMs)	Trichostatin A
Adipocytokine signaling	Prochlorperazine, fluphenazine, rosiglitazone, troglitazone
PC3 cell line panel	
Steroid hormone biosynthesis	Thalidomide, nifenazone, isotretinoin
Antigen processing and presentation	Ebselen, scoulerine, nifedipine, disulfiram, 15-delta prostaglandin J2, withaferin A, thiostrepton, mometasone, MG-262, phenoxybenzamine, 1,4-chrysenequinone
PPAR signaling pathway	Pioglitazone, gliquidone, mifepristone, glibenclamide, rosiglitazone, troglitazone
MCF7 cell line panel	
Tryptophan metabolism Leukocyte transendothelial migration Toll-like receptor signaling NOD-like receptor signaling	Anisomycin, cephaeline
Arachidonic acid metabolism	Isotretinoin
Retinol metabolism	Tretinoin
Antigen processing and presentation	Tanespimycin, alvespimycin, geldanamycin, disulfiram, thiostrepton, 15-delta prostaglandin J2, ebselen, phenoxybenzamine, piperlongumine, MG-262
Metabolism of xenobiotics by cytochrome P450 Steroid hormone biosynthesis	Lansoprazole, tetrahydroalstonine, oxyphenbutazone, omeprazole, nifuroxazide, calycanthine . . .

Subramanian *et al.*, 2005), using the HL60, PC3 and MCF data set. The R code for GSEA analysis was downloaded from the following website http://www.broadinstitute.org/gsea/downloads.jsp. For each treatment within the same cell line panel (as shown in Table 3), we first ranked all the genes according to the treatment/control ratio. Then for each KEGG pathway, a Kolmogorov–Smirnov running sum is computed. The resulting enrichment score (ES, real number between −1 and +1) indicates how significantly a specific pathway is up-regulated (ES > 0) or down-regulated (ES < 0) upon this treatment. We then compared the Precision–Recall curves generated from |ES| by GSEA with those generated from matrix |**X**| by FacPad, still using the CTD as the validation resource (Figure 6). Precision–Recall curves are more suitable to the current situation than the ROC curves (True Positive Rate versus False Positive Rate). Comparing the different metrics used by Precision–Recall curve and ROC curve:

$$precision = \frac{TP}{TP + FP},$$

$$recall = true\_positive\_rate = \frac{TP}{TP + FN},$$

$$false\_positive\_rate = \frac{FP}{FP + TN}$$

As a result of the incompleteness of drug–pathway interactions documented in CTD (the comprehensive list of actual drug–

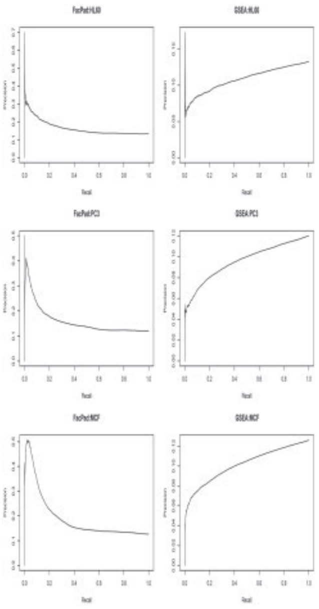


Fig. 6. Precision–Recall curves of pathway–drug associations estimated by FacPad versus those by GSEA, using CTD as the validation set

pathway interactions is unknown and it is impossible to get the accurate validation set), the calculation for *FP* (false positives) and *TN* (true negatives) is much less reliable than *TP* (true positives) and *FN* (false negatives). Therefore, Precision–Recall curves are more appropriate and less misleading than ROC curves when only a small proportion of all the positives are known in the validation set, which has been noted in protein–protein interaction prediction studies (Hue *et al*., 2010) as well as in the machine learning community (Davis and Goadrich, 2006). From Figure 6, it can be seen that FacPad always gives a better Precision–Recall curve than GSEA. The different performance may result from the fact that GSEA is a rank-based method and treats all the genes in the same pathway with equal weight. In contrast, FacPad assigns different loadings to genes in the same pathway and borrows information across all the treatments.

3.5 A further note on the choice of sparsity structure of loading matrix W

For the model presented here, the sparsity structure of loading matrix W is completely determined by the pathway–gene association relationship derived from KEGG database. However, the loading matrix W may be sparser in that some gene–pathway coefficients might be close to 0 even if their associations are documented in KEGG. Indeed, when we check on the estimated loadings of the genes in the same pathway, it is usually the case that only a small proportion of genes have high loadings (e.g. with a absolute value >0.5). One of the advantages of the Bayesian sparse factor framework utilized by FacPad is that it allows convenient incorporation of more flexibility in the sparsity structure of the loading matrix W. To achieve this, we just have to add another layer between loading matrix W and the binary gene–pathway association matrix L. It is another binary matrix Z representing the true sparsity structure of matrix W. Each entry in matrix Z is assumed to follow a Bernoulli distribution with the probability of taking 1/0 depending on the corresponding entry in matrix L:

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

$$W_{g,k}=\begin{cases}0, & \mathit{if} \quad Z_{g,k}=0\\ N(0,\tau_w^{-1}), & \mathit{if} \quad Z_{g,k}=1\end{cases}$$

As for the collapsed Gibbs sampling, it is only needed to add one more step for the inference of matrix Z before the inference of matrix W (see Appendix 4 of Supplementary Material for details). The parameter η_0 and η_1 are used to fine tune the similarity between the link matrix L and the true sparsity structure matrix Z. If both η_0 and η_1 are set to 0, the model goes back to the original version described earlier.

We compared the result of the modified model with the original simplified version using the HL60 data set. We set η_0 to 0 and tested five different values of η_1: 0, 0.2, 0.4, 0.6 and 0.8. We ran the Collapsed Gibbs sampling algorithm for 20 000 iterations and recorded the result every other 10th iteration. The first 10 000 iterations were discarded as burn-in period, whereas the remaining samples were averaged to estimate the binary indicator matrix Z, the absolute value of loading matrix $|W|$ and the latent factor activity matrix $|X|$. As shown in Table 6, the total number of 1-entries in matrix Z decreases with η_1 increasing, but the 1-entries inferred using $\eta_1=0.2$, 0.4, 0.6, 0.8 are always a subset of the 1-entries inferred using $\eta_1=0$. The corresponding fold enrichment is shown in parentheses, calculated as: $N_{1\cap 2}\times \mathit{length}(Z)/(N_1\times N_2)$. Figure 7 shows the distribution of non-zero loadings in matrix $|\mathbf{W}|$ (as indicated by estimated matrix Z) for different η_1 values used. Large η_1 leads to fewer but higher non-zero loadings.

Table 6. Number of 1-entries in matrix Z estimated using different η_1

H_1	0	0.2	0.4	0.6	0.8
0	4466	2692 (46.75)	2411 (46.75)	2253 (46.75)	2140 (46.75)
0.2		2692	2247 (72.28)	2171 (74.73)	2008 (72.77)
0.4			2411	2104 (80.87)	2023 (81.86)
0.6				2253	1979 (85.70)
0.8					2140

The diagonal shows the result using a specific η_1 and the off-diagonal elements shows the number of common 1-entries estimated using two different η_1.

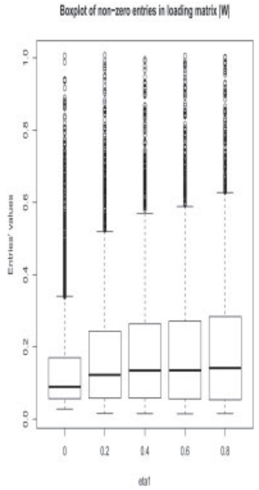


Fig. 7. Boxplot of non-zero loading in matrix $|\boldsymbol{W}|$

4 DISCUSSION

In this article, we have presented a Bayesian sparse factor analysis model, FacPad, for the inference of drug–pathway associations using treatment response data generated from microarray platforms. This method treats biological pathways as latent factors and explains the gene-wise treatment response variation in items of the latent pathway variables. This approach is different from previous methods for such data, since it does not merely calculate the similarity among different drugs, but tries to detect the association strength between drugs and pathways through large-scale data mining. Although this method was designed for the analysis of gene expression data, it can also be directly applied to proteomic and other high-dimensional data sets. The latter half of this article focuses on information mining on the

Connectivity Map data, which represents an excellent repository of treatment response profiles of a broad range of small bioactive compounds. We have shown that FacPad has good performance for the inference of true positive drug–pathway associations. It can also predict the target pathways and possible mechanism of action for drugs with little known information in existing drug databases. We implemented the model as an R package for public access. It is worthy of note that the FacPad model was developed under the assumption that all the latent factors needed for decomposition of the data matrix are known (here, the latent factors are KEGG pathways mapped to by the measured genes). Therefore, it should not be applied directly without modification when the real latent factors are completely unknown. Similar to other Bayesian method, relatively good computational resource is required for running this program on large-scale data sets. We used the Yale computing cluster which usually takes 18 h to finish 5000 iterations of Gibbs sampling on the HL60 data set. For future studies, it will be of interest to try different encoding of the latent factors. Instead of using a linear factor model, other data mining techniques may also achieve good result.

ACKNOWLEDGEMENT

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

Funding: National Institutes of Health (GM59507 to H.Z. and NIH R21-GM084008 to Ning Sun). (NSF DMS 1106738 to H.Z.).

Conflict of Interest: none declared.

REFERENCES

Bader,G.D. *et al.* (2006) Pathguide: a Pathway Resource List. *Nucleic Acids Res.*, **34**, D504–D506.

Berti,E. *et al.* (1998) Mometasone furoate decreases adhesion molecule expression in psoriasis. *Eur. J. Dermatol.*, **8**, 421–426.

Brettschneider,J. *et al.* (2008) Quality assessment for short oligonucleotide microarray data. *Technometrics*, **50**, 241–264.

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

Caspi,R. *et al.* (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **40**, D742–D753.

Davis,A.P. *et al.* (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.*, **39**, D1067–D1072.

Davis,J. and Goadrich,M. (2006) The relationship between precision–recall and ROC curves. *ICML ‘06 Proc.* 23^{rd} *Int. Conf. Machine Learn*, 233–240.

He,Z. *et al.* (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One*, **5**, e9603.

Huang da,W. *et al.* (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Huang da,W. *et al.* (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Hue,M. *et al.* (2010) Large-scale prediction of protein–protein interactions from structures. *BMC Bioinformatics*, **11**, 144.

Iorio,F. *et al.* (2009) Identifying network of drug mode of action by gene expression profiling. *J. Comput. Biol.*, **16**, 241–251.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. USA.*, **107**, 14621–14626.

Iskar,M. *et al.* (2010) Drug-induced regulation of target expression. *PloS Comput. Biol.*, **6**, e1000925.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kitchen,D.B. *et al.* (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, **3**, 935–949.

Kotera,M. *et al.* (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.*, **802**, 19–39.

Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer*, **7**, 54–60.

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

Lane,M.A. and Bailey,S.J. (2005) Role of retinoid signalling in the adult brain. *Prog. Neurobiol.*, **75**, 275–293.

Luthi-Carter,R. *et al.* (2000) Decreased expression of striatal signaling genes in a mouse model of Huntington’s disease. *Hum. Mol. Genet.*, **9**, 1259–1271.

Ma,H. and Zhao,H. (2012) iFad: an integrative factor analysis model for drug-pathway association inference. *Bioinformatics*, **28**, 1911–8.

Mootha,V.K. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Nagamine,N. *et al.* (2009) Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. *PLoS Comput. Biol.*, **5**, e1000397.

O’Reilly,K.C. *et al.* (2006) Chronic administration of 13-cis-retinoic acid increases depression-related behavior in mice. *Neuropsychopharmacology*, **31**, 1919–1927.

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

Rattray,M. *et al.* (2009) Inference algorithms and learning theory for Bayesian sparse factor analysis. *J. Phys.: Conf. Ser.* **197**, 012002 doi:10.1088/1742-6596/197/1/012002.

Sabatti,C. and James,G.M. (2006) Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, **22**, 739–746.

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

Smalley,J.L. *et al.* (2010) Application of connectivity mapping in predictive toxicology based on gene-expression similarity. *Toxicology*, **268**, 143–146.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA.*, **102**, 15545–15550.

Tsumuro,T. *et al.* (2005) Effects of mometasone furoate on a rat allergic rhinitis model. *Eur. J. Pharmacol.*, **524**, 155–158.

Vina,D. *et al.* (2009) Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. *Mol. Pharm.*, **6**, 825–835.

West,M. (2003) Bayesian factor regression models in the ‘Large p, Small n’ paradigm. *Bayesian Stat.*, **7**, 733–742.

Xie,L. *et al.* (2012) Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annu. Rev. Pharmacol. Toxicol.*, **52**, 361–379.

Yamanishi,Y. *et al.* (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.

Yamanishi,Y. *et al.* (2010) Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, **26**, i246–i254.

Young,D.W. *et al.* (2008) Integrating high-content screening and ligand–target prediction to identify mechanism of action. *Nat. Chem. Biol.*, **4**, 59–68.

Zhao,S. and Iyengar,R. (2012) Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. *Annu. Rev. Pharmacol. Toxicol.*, **52**, 505–521.

Zhou,T. *et al.* (2010) Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. USA.*, **107**, 4511–4515.