RESEARCH ARTICLE

A novel method of using Deep Belief Networks and genetic perturbation data to search for yeast signaling pathways

Songjian Lu[1]*, **Xiaonan Fan**[1,2], **Lujia Chen**[1], **Xinghua Lu**[1]

1 Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **2** Department of Automation, Northwestern Polytechnical University, Shanxi, People's Republic of China

* songjian@pitt.edu

OPEN ACCESS

Citation: Lu S, Fan X, Chen L, Lu X (2018) A novel method of using Deep Belief Networks and genetic perturbation data to search for yeast signaling pathways. PLoS ONE 13(9): e0203871. https://doi.org/10.1371/journal.pone.0203871

Editor: Aamir Ahmed, King's College London, UNITED KINGDOM

Received: January 24, 2018

Accepted: August 29, 2018

Published: September 12, 2018

Data Availability Statement: Data are from the Kemmeren's study, which can be found at Gene Expression Omnibus (GEO) with accession number GSE42527.

Funding: This work was supported by U.S. National Library of Medicine (Grant R00LM011673 to Dr. Songjian Lu) and National Human Genome Research Institute (Grant U54HG008540 to Xinghua Lu). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Perturbing a signaling system with a serial of single gene deletions and then observing corresponding expression changes in model organisms, such as yeast, is an important and widely used experimental technique for studying signaling pathways. People have developed different computational methods to analyze the perturbation data from gene deletion experiments for exploring the signaling pathways. The most popular methods/techniques include K-means clustering and hierarchical clustering techniques, or combining the expression data with knowledge, such as protein-protein interactions (PPIs) or gene ontology (GO), to search for new pathways. However, these methods neither consider nor fully utilize the intrinsic relation between the perturbation of a pathway and expression changes of genes regulated by the pathway, which served as the main motivation for developing a new computational method in this study. In our new model, we first find gene transcriptomic modules such that genes in each module are highly likely to be regulated by a common signal. We then use the expression status of those modules as readouts of pathway perturbations to search for up-stream pathways. Systematic evaluation, such as through gene ontology enrichment analysis, has provided evidence that genes in each transcriptomic module are highly likely to be regulated by a common signal. The PPI density analysis and literature search revealed that our new perturbation modules are functionally coherent. For example, the literature search revealed that 9 genes in one of our perturbation module are related to cell cycle and all 10 genes in another perturbation module are related by DNA damage, with much evidence from the literature coming from *in vitro* or/and *in vivo* verifications. Hence, utilizing the intrinsic relation between the perturbation of a pathway and the expression changes of genes regulated by the pathway is a useful method of searching for signaling pathways using genetic perturbation data. This model would also be suitable for analyzing drug experiment data, such as the CMap data, for finding drugs that perturb the same pathways.

Competing interests: The authors have declared that no competing interests exist.

Background

Understanding cellular signaling pathway systems is one of the major tasks those in the systems biology field undertake [1]. Many important cell activities, such as proliferation and apoptosis, can be regulated by signaling pathways that accept signals from the surface of cells, where the pathways regulate cell activities by adjusting the expression levels of corresponding down-stream genes. Hence, the study of pathways can help us to understand the mechanism of diseases, such as cancer, that are caused by genetic problems [2, 3].

One well established technology that can be used to study the cell signaling system is genetic perturbation experiments, i.e., observing cell expression profile changes by deleting protein-coding genes in model organisms, such as yeast. For example, Hughes et al. performed a pioneering study of yeast (*Saccharomyces cerevisiae*) signaling systems by generating and studying genome-wide mRNA expression profiles with the deletion of 276 protein-coding genes [4]. Very recently, Kemmeren et al. generated a new data set with the mRNA expression profiles of 1484 deletion mutations of protein-coding genes for the study of yeast regulatory systems [5]. This type of experiment has generated a large amount of expression data [4–6] that provides opportunities for studying the signaling system using computational methods.

One group of popular computational methods is clustering based, such as the hierarchical or k-means clustering. The basic idea is that if two genes have similar expression profiles across all samples or the deletions of two genes have similar genome-wide expression profiles, then these two genes are functionally related. Kemmeren et al. used hierarchical clustering to study the expression data and found that if genes are in the sample protein complex or the same pathway, then genome-wide expression profiles of deletions of these genes were significantly similar [5]. There are some other works [7–11] that combined expression data with other knowledge or techniques to search for or study signaling pathways. For example, Steffen et al. combined gene expression data, protein-protein interaction network, and k-mean algorithm to search for sub-network [9]. Their basic idea was that genes in a sub-network were more likely to belong to a pathway if they were in one cluster obtained from clustering the expression data. Zhao et al. applied expression profile and mutual exclusivity to find pathways related to cancer development [10]. They thought that if mutations of genes were mutually exclusive among tumors, and furthermore, gene expressions of those genes were also similar across all tumors, then those genes were likely to be on the same pathway. In a summary, the purpose of using expression data in previous works was similar, i.e., genes in a pathway should have similar expression profiles across all samples or expressed genes.

Our major motivation for proposing a new computational model to search for signaling pathways in this work is that (to our best knowledge) previous methods did not consider or not fully utilize the intrinsic relation between the perturbation of a pathway and expression changes of genes regulated by the pathway. It is obvious that if a deletion perturbs a pathway, i.e., a gene/protein in a pathway has been deleted, then expression levels of genes regulated by the pathway should change significantly. However, when we measure those expression changes, certain random perturbations, such as the variance of the microarray products and the impact of artificial factors in the experiments, are hard to avoid. For example, researchers may have noticed expression differences of genes among wild-type or control samples. People may also have found that in many data sets, even the deletion of the same gene in two samples under the same condition, the significantly changed genes in the two samples were quite different. So in the expression data resulted from the deletion of a gene in a pathway, besides the genes regulated by the pathway, some other random genes may also be differentially expressed. If we used genome-wide expression profiles to study the relations between deleted genes using some traditional methods, such as the hierarchical or k-means clustering, those random perturbations would cause problems as expression of a large number of genes not regulated by the pathway also make contribution in those clustering methods. In this project, in order to reduce above problems, we first search for gene modules, called transcriptomic modules, such that genes in each transcriptomic module are highly likely to be regulated by a common pathway. We then use the expression status of each transcriptomic module as a readout of pathway perturbations to search for an up-stream perturbation module such that genes in the perturbation module come from the same pathway. *The random perturbations should be greatly reduced if we only consider the expression changes of genes regulated by each pathway respectively when we search for up-stream perturbation modules.*

We use a deep learning technique called deep belief network (DNB) [12, 13] to search for transcriptomic modules as it can learn the hierarchical structure that exists within the differentially expressed genes of the perturbation data. The DBN is a machine learning technique that was originally developed for image processing, such as face recognition. A DBN may have one visible layer and multiple hidden layers. When a DBN is applied for face recognition, its nodes in the first, second, and third hidden layers can group pixels that make edges (line, curve segments, etc.), components (eye, nose, mouse etc.), and faces together, respectively [14, 15], i.e., it can learn the hierarchical structure that exists within the input data. In the cascade structure of a pathway system, a gene/protein in up-stream usually controls more genes (gene expressions) than a gene in down-stream does. However genes controlled by different genes along the same pathway should have a hierarchical structure. We use DBN to learn this hierarchical structure and to search for down-stream transcriptomic modules such that genes in each transcriptomic module are commonly regulated by one signaling pathway. The major difference of using DBN and other clustering methods, such as hierarchical clustering or k-means method, is that the DBN is finding gene modules such that genes in each module are co-differentially expressed in a number of samples that are statistically significant while the hierarchical clustering and k-means method are searching for genes with similar expression values in all samples. So for the hierarchical clustering and k-means method, expression values of genes in samples that do not have the significant expression changes of those genes also affect the clustering results. As the lengths of pathways are usually not very long, in the perturbation data set, genes regulated by each signaling pathway should not be co-differentially expressed in many samples. Using DBN is a better way to find genes regulated by each pathway as genes regulated by each pathway were only co-differentially expressed in a very small number of perturbed samples.

Our paper is organized as follows. After the introduction section, we introduce the methods of our model in detail. We then introduce the results, including evaluation. Finally, we present our conclusions.

Methods

Data collection and preprocessing

We collected the gene expression data of 1484 samples [5], where each sample is the mRNA expression profile of the deletion of one protein-coding gene in *Saccharomyces cerevisiae*. Each profile includes expression level in the form of standard deviation, average transcription level changes (fold changes) in the mutant relative to 428 WTs, and *p*-values. In the preprocessing step, we used the setting of the paper [5] to find what genes were differently expressed under the deletion, i.e. a gene was considered to be differently expressed if its fold change was at least 1.7 and the *p*-value was less than 0.05. After the preprocessing, we obtained a 0/1 matrix such that each row is for a measured gene and each column for a sample (perturbed gene); a value 1 represented that a gene was differently expressed in a sample; otherwise the value was set to 0.

This 0/1 matrix was used to train the Deep Belief Network (DBN). The 0/1 matrix depends on the threshold setting, which may lead to the change of trained DBN. However, if a method is stable, the results should not change too much for a little change of threshold setting. We have tested to use a fold change cutoff of 1.75 and 1.65 to make 0/1 matrices, respectively. We found that the results of DBN from 0/1 matrices of different fold change cutoffs were very similar, which provides evidence that the DBN model is quite s. Note: when the DBN program loads this 0/1 matrix, the number of nodes in the visible layer is set as the number of measured genes in the 0/1 matrix.

Training the DBN with the 0/1 matrix

In the introduction, we stated that differentially expressed genes caused from the perturbations of different locations of a signaling pathway have a hierarchy structure that can be discovered by a DBN. People usually use three or four hidden layers for DBN, where the default number of hidden layers is four in the Matlab codes provided by Hinton et al. In this work, we used a DBN with four hidden layers to learn this hierarchy structure, where the number of nodes in the visible layer is the number of measured genes (row number of the 0/1 matrix), and the number of nodes in the first, second, third and fourth hidden layers are 217, 160, 94, and 166 respectively. To obtain a good performance, setting a proper number of nodes in each hidden layer is important. The number of nodes needed in each hidden layer depends on the input training data. As there is no "gold standard", people usually adjust those parameters manually. In this project, we introduce a better way to estimate the number of nodes needed in each hidden layer. After the DBN is trained by a given input data, for each node in any hidden layer and each sample in the input data, the DBN returns the probability that the hidden node is activated in the sample. Hence, we can know how many nodes in each hidden layer have been activated with a probability of p in at least one sample. We found that for a fixed probability p, such as 0.75, if we set the number of nodes in all hidden layers to k and gradually increased this k, the number of nodes that were activated with probability p in at least one sample in each layer fluctuated around a certain number (refer to Fig 1). For example, if we set k to be 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, and 350, then the number of nodes that were activated with a probability of 0.75 at the first hidden layer would be 100, 125, 150, 175, 189, 193, 191, 198, 217, 191, and 214 respectively. We found that the DNB was able to achieve good performance if we found the maximum number obtained from the different settings of k for each hidden layer and then used this maximum number to set this hidden layer, for example, in the previous case, setting the number of nodes in the first hidden layer to be 217.

Obtaining the transcriptomic modules

After the DBN was trained, for each node T in the first hidden layer, we learned the weights for all edges from the node T to all nodes (genes) in the visible layer. Each edge weight represents how strongly the value of the visible layer is affected by the value of node T. We computed the mean μ and standard deviation δ from all the weights of the edges from T to all of the nodes in the visible layer. Then we used μ and δ to set a threshold, a p-value of 0.05, for choosing a set of nodes (genes) S in the visible layer and considered them to be regulated by the node T. We considered genes in S to be a transcriptomic module.

Searching for up-stream perturbation modules

For each transcriptomic module, we tried to find a set of genes that was highly likely to be on the pathway regulating the expression of the transcriptomic module. We called this set of genes a perturbation module, as the deletion of any gene in the perturbation module would

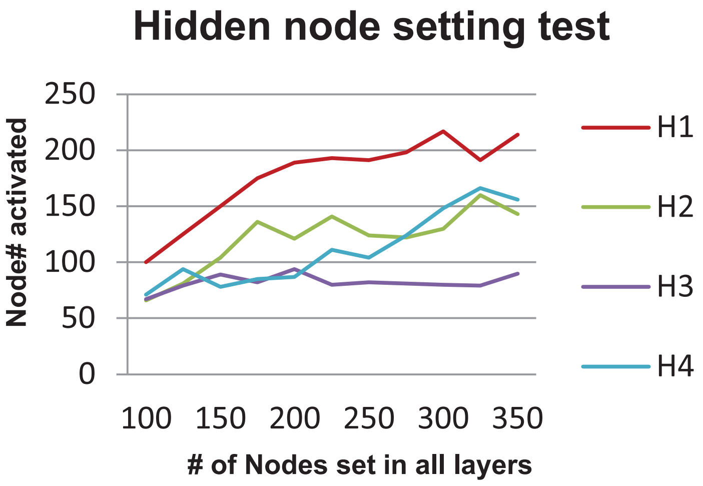

Fig 1. Numbers of hidden nodes that were set and actually activated in all hidden layers.

https://doi.org/10.1371/journal.pone.0203871.g001

perturb the expression of the corresponding transcriptomic module. We only considered genes that were deleted in the data set we collected. For each down-steam transcriptomic module, we **first** found a gene, *called the initial gene*, such that the deletion of the initial gene caused the most number of genes in the transcriptomic module to be differently expressed. We **then** iteratively added new genes into the solution such that every time a new gene was added, it would have the shortest average distance to all of the other genes in the previous solution. For a transcriptomic module, the distance between any two genes G_1 and G_2 was decided by the *normalized* expression level changes (normalized fold changes) of the genes in the transcriptomic module with the deletions of G_1 or G_2, where the normalized fold change is defined as:

$$norm(x) = \begin{cases} 2, & x \geq 1.7 \text{ and } p_{\_\text{Value}} < 0.05 \\ 1.7, & x \geq 1.7 \text{ and } p_{\_\text{Value}} \geq 0.05 \\ x, & -1.7 < x < 1.7 \\ -1.7, & x \leq -1.7 \text{ and } p_{\_\text{Value}} \geq 0.05 \\ -2, & x \leq -1.7 \text{ and } p_{\_\text{Value}} < 0.05 \end{cases}.$$

We normalized the fold changes of the gene expression levels as we wanted the fold change on a gene not to contribute to the distance if it is up- or down-regulated significantly in both gene deletions. As in the original data, the expression of a gene was considered to change significantly if the fold change was at least 1.7 and the p-value was less than 0.05, so we wanted to distinguish cases of p-value ≥ 0.05 and p-value<0.05. When the fold change is at least 1.7, we set the value to be ±2 when the p-value<0.05. Specifically, suppose that the transcriptomic module has genes $g_1, g_2, \ldots, g_t$; the deletion of G_1 would cause the expression changes of those t genes to be $u_1, u_2, \ldots, u_t$, and the deletion of G_2 would cause the expression change of those t genes to be $v_1, v_2, \ldots, v_t$; then the distance of G_1 and G_2 would be the minimum Euclidean distances between vectors $(u_1, u_2, \ldots, u_t)$ and $(v_1, v_2, \ldots, v_t)$ and the Euclidean distances between vectors $(u_1, u_2, \ldots, u_t)$ and -$(v_1, v_2, \ldots, v_t)$. We consider the Euclidean distances between vectors $(u_1, u_2, \ldots, u_t)$ and -$(v_1, v_2, \ldots, v_t)$ as the deletion of a gene on the pathway may inhibit the signal while the deletion of another gene on the pathway may enhance the signal. Therefore, the expression changes of genes regulated by a pathway may be in the reverse direction for the deletion of different genes on the same pathway.

Results

We obtained 217 transcriptomic modules. Correspondingly, we found 217 up-stream perturbation modules. In this section, we first give a systematic evaluation of the transcriptomic and perturbation modules. We then present more detail of some of our perturbation modules.

In our hypothesis, genes in a transcriptomic module are highly likely to be regulated by a pathway or even by a transcription factor. Hence, to evaluate our transcriptomic modules, we first determined whether genes in a transcriptomic module were enriched in genes regulated by known transcription factors. Then we investigated whether the genes in the transcriptomic modules were functionally coherent.

Verifying transcriptomic modules with transcription factors

In the gene expression data that we collected, there existed deletions of 67 transcription factors that caused at least 10 genes to be differently expressed. As a result, we could obtain genes that are regulated by those 67 transcription factors. Using enrichment analysis, we first checked the overlap of genes in the transcriptomic modules with genes regulated by those 67 transcription factors. Remember that genes in each transcriptomic module are regulated by a node in the first hidden layer. After the DBN was trained using given training data, we obtained information about the probability that a node in the first hidden layer would be activated in each sample in the training data. We chose the 0.95 as the confidence threshold to decide if a node in the first hidden layer would be activated in a sample. Hence, we obtained information about how many times a node in the first hidden layer was activated in the training data.

By intuition, we know that if a node *h* is activated in only a very few or even no samples, then the weights from the node *h* to all nodes in the visible layer will not be well trained. Therefore, genes in transcriptomic modules regulated by the node *h* should be less reliable than genes in transcriptomic modules regulated by a node that is activated many times. Our results supported this hypothesis. We split our transcriptomic modules into three groups according to the number of activations of their corresponding nodes in the first hidden layer, i.e. modules in group 1, 2, and 3 are regulated by nodes that are activated 0, between 1 and 30, and more than 30 times, respectively. Our results show that the average enrichment *p*-values (negative log value with base 2) for transcriptomic modules in group 1, 2, and 3 were 22.03, 42.21, and 101.2, respectively (refer to Fig 2). Looking back to the original space, on average, the enrichment *p*-values for transcriptomic modules in group 2 were 9.8×10^5 fold better than those for transcriptomic modules in group 1, and the enrichment *p*-values for transcriptomic modules in group 3 were 5.7×10^{17} fold better than those for transcriptomic modules in group 2.

Verifying transcriptomic modules using Gene Ontology (GO)

We also verified our transcriptomic modules using Gene Ontology to determine whether the genes in each transcriptomic modules were functionally coherent. As we supposed that the genes in each transcriptomic module are regulated by a signaling pathway, they should be functionally coherent. For each transcriptomic module, we searched for a GO term such that genes in the transcriptomic modules were most enriched in the GO term, i.e. had the minimum *p*-value for the hypergeometric test. We still compared the enrichment *p*-values (negative log value with base 2) for the transcriptomic modules in the three groups above. These results also showed that the transcriptomic modules in group 3 had the best enrichment *p*-values, where the average *p*-values (in log space) for the transcriptomic modules in group 1, 2, and 3 are 14.38, 17.36, and 24.02, respectively (refer to Fig 3). A global enrichment analysis of top 20 most enriched GO terms and their matched transcriptomic modules can be found at S1 Fig

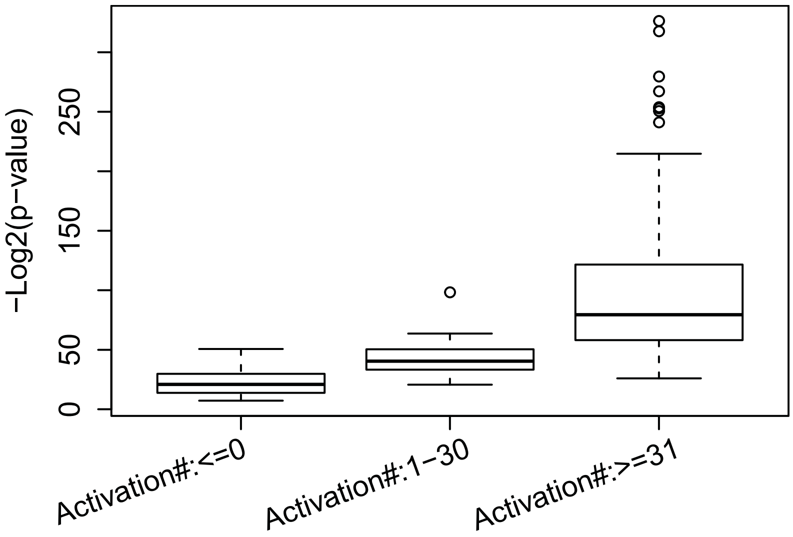

Fig 2. Comparing the TF enrichment analysis of genes regulated by the first layer hidden nodes with different activation numbers.

https://doi.org/10.1371/journal.pone.0203871.g002

The result shows that many top 20 enriched GO terms are related to metabolic process and some transcriptomic modules can be significantly enriched in more than one GO term.

Verifying perturbation modules using protein-protein interactions (PPIs)

According to our hypothesis, the genes in each perturbation module are highly likely to be on the same signaling pathway. As PPIs are important for signal transduction [16, 17], we expected that there would be more PPIs among genes in our perturbation modules than among genes that are obtained using a random process. As the lengths of signaling pathways are usually not very long, the sizes of perturbation modules should not be too large. To test our

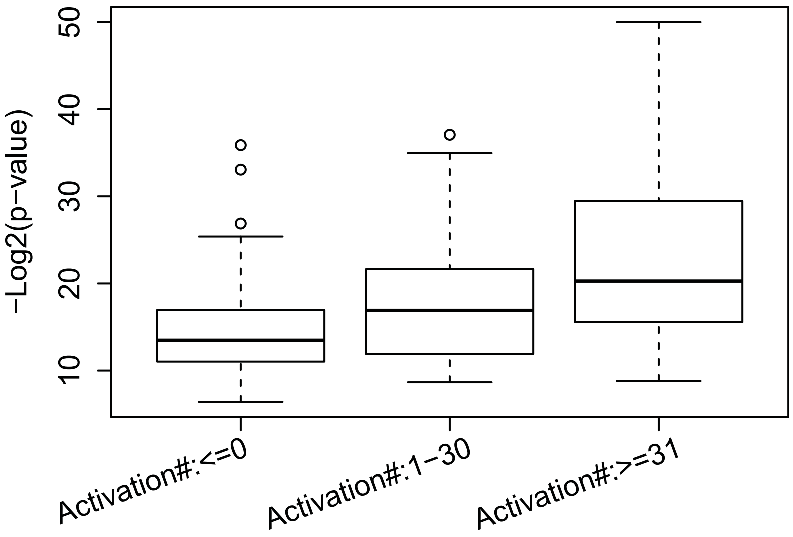

Fig 3. Comparing the GO enrichment analysis of genes regulated by the first layer hidden nodes with different activation numbers.

https://doi.org/10.1371/journal.pone.0203871.g003

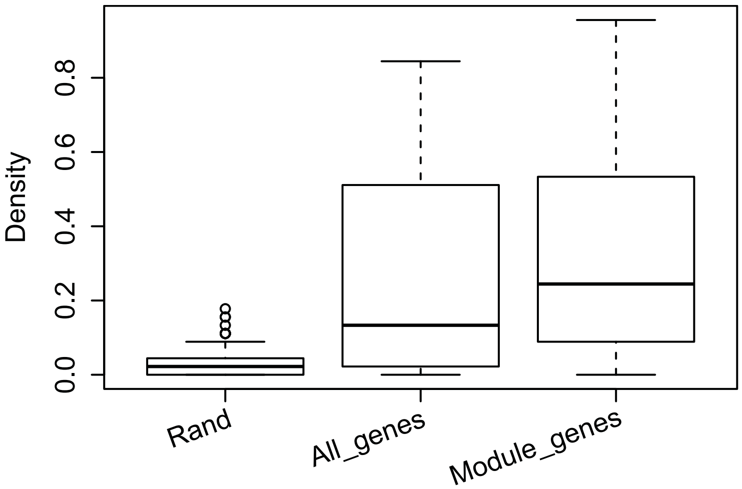

Fig 4. Comparing the PPI density of genes in perturbation modules obtained from different methods.

https://doi.org/10.1371/journal.pone.0203871.g004

model, we fixed the size of the perturbation modules to 10. The results agree with our expectation (refer to Fig 4). We found that the average PPI density, i.e. the ratio of the actual PPI number and total number all gene pairs, for 10 genes randomly chosen from the deleted genes in the expression data set was 0.034 (marked as "Rand") while the average density for genes from our perturbation module was 0.33 (Marked as "Module_genes"). If we use the same algorithm for finding our perturbation modules, but instead of constraining genes in the transcriptomic modules we use all genes to search for 10 genes, the average PPI density is only 0.27. Hence, by using only the genes in the transcriptomic modules to search for perturbation modules, we can greatly improve performance in terms of PPI density.

Verifying perturbation modules using Gene Ontology (GO)

We also verified our perturbation modules using Gene Ontology and found that results of the analysis of our perturbation was similar to those of the analysis using PPI density. The average enrichment *p*-values (negative log value with base 2) of random perturbation modules was 9.48 while the average enrichment p-values of our perturbation modules was 25.50, where the difference was 6.65×10^4 in the original space. In the sample time, the average enrichment for perturbation modules obtained from all genes was 18.76 (refer to Fig 5). Hence, the GO analysis also proved that transcriptomic modules could be greatly helpful in finding perturbation modules. A global enrichment analysis of top 20 most enriched GO terms and their matched perturbation modules can be found at S2 Fig GO enrichment analysis also shows that the normalization of fold change also improve the performance of the perturbation module finding (refer to S3 Fig).

Literature search revealed that genes in our perturbation modules are functionally coherent

We conducted a literature search to study genes in our perturbation modules and found that the genes in our perturbation modules are functionally coherent. As the literature search has to be done manually, it is hard to verify many modules. In this section, we only report the results of literature search for two perturbation modules. One perturbation module, denoted as Module-30, has genes BIM1, JNM1, MMS22, NPL3, RAD18, RAD50, RAD52, RMI1, SGS1,

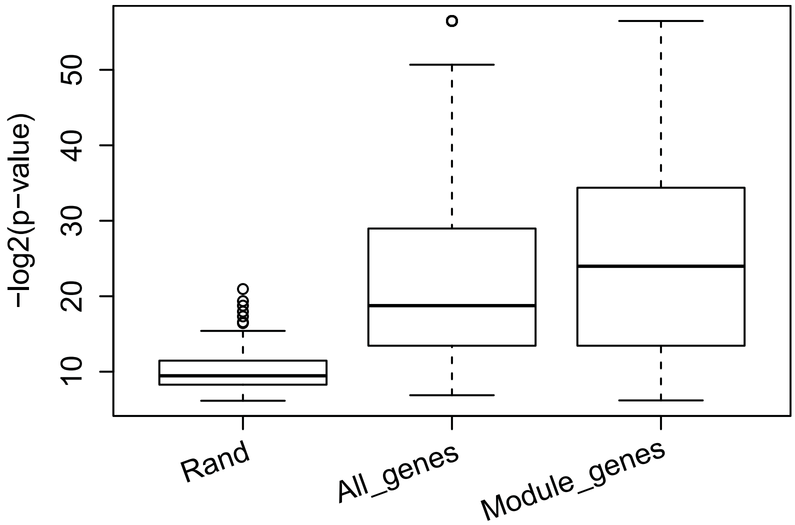

Fig 5. Comparing the GO enrichment analysis of genes in perturbation modules obtained using different methods.

https://doi.org/10.1371/journal.pone.0203871.g005

and TOP3. There exist many protein-protein interactions among these10 genes (refer to Fig 6). We found that 9 of these 10 genes are associated with the cell cycle. Voncken et al. found that BIM1 is cell cycle-regulated and associated with the G(1)-phase of the cell cycle [18]. McMillan [19] and Wang [20] et al. revealed that JNM1 regulates the spindle orientation during the mitotic cell cycle. Vaisica et al. showed that the deletion of MMS22 caused an abnormal cell cycle [21]. Dovey et al. also verified that the loss of MMS22 had an impact on the S- and G2-phases of the cell cycle [22]. Bi et al. showed that RAD18 regulates the recovery from S-phase checkpoint-mediated arrest [23]. Zhu [24] and Gatei et al. [25] found that RAD50 is related to cell cycle regulation. Lisby discovered that RAD50 is associated with the DNA repair and recombination centers during the S-phase of the cell cycle [26]. Xu et al. presented that RMI1 is related to the M-phase of the cell cycle [27]. Balogun et al. found that the loss of SGS1 significantly impairs activation of cell cycle arrest [28]. Mankouri et al. showed that TOP3 is required for normal S-phase progression after DNA damage [29]. Therefore, it is highly likely that genes in Module-30 regulate the cell cycle.

In another of our perturbation modules, denoted as Module-80 (refer to Fig 7), a literature search showed strong evidence that this perturbation is associated with the function/pathway related to DNA damage as all 10 genes in the Module-80 have been proven to be related to DNA damage. For example, Sharp [30] and Hu [31] et al. found that ASF1 is related to DNA damage. Clausing et al. showed that BUR2 is associated with functions of DNA repair [32]. Fumasoni et al. presented that the DNA damage tolerance relies on CTF4 [33]. Crabbé et al. indicated that CTF18 is essential for DNA damage control [34]. Dovey et al. verified that loss of MMS22 results in the accumulation of spontaneous DNA damage. Xu et al. presented that MRC1 is required for DNA damage checkpoint activation [35]. Karras et al. found the regulation of the RAD6 pathway to DNA damage [36]. Hedglin et al. studied RAD6 activity related to DNA damage tolerance [37]. Chahwan [38] and Roset [39] et al. presented an association of RAD50 to DNA damage. Sidorova and Breeden showed that SWI6 has a function in response to DNA damage [40]. Mankouri [29] and Mohanty [41] et al. presented that TOP3 is related to DNA damage. It is observed that the three genes MMS22, RAD50, and TOP3 of Module-80 are also in Module-30, which mainly regulates the cell cycle. Hence, it is very likely that these 3

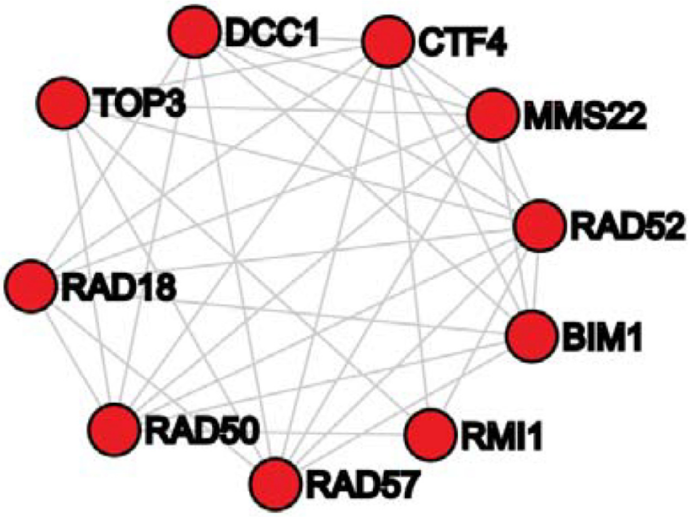

Fig 6. Protein-protein interaction subnetwork of genes in the perturbation module-30.

https://doi.org/10.1371/journal.pone.0203871.g006

genes play multiple roles and genes in Module-80 are in a pathway that regulates a partial function of cell cycle–DNA damage.

We went back to check transcriptomic modules and found that these two perturbation modules and their corresponding transcriptomic modules were functionally coherent. Gene

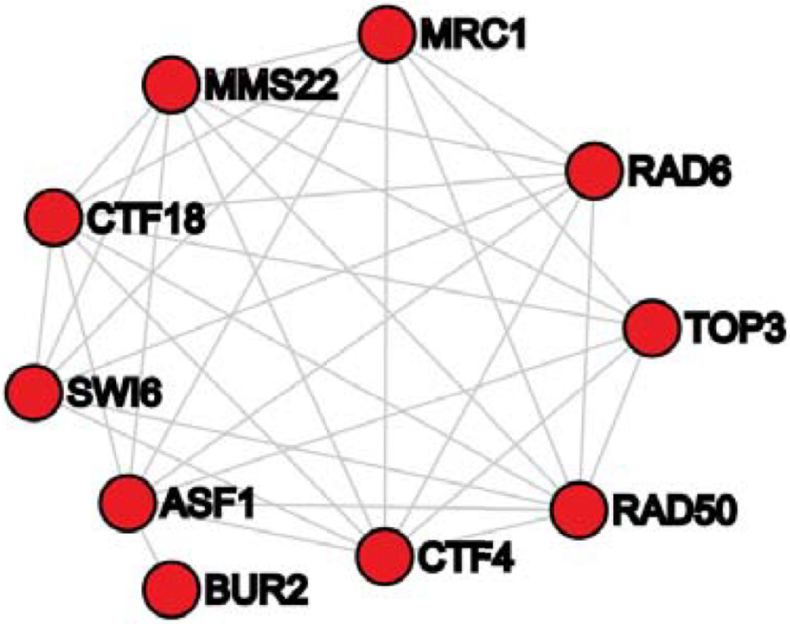

Fig 7. Protein-protein interaction subnetwork of genes in the perturbation module-80.

https://doi.org/10.1371/journal.pone.0203871.g007

Ontology analysis showed that many genes in the transcriptomic module corresponding to Module-30 are associated with the cell cycle process while many genes in the transcriptomic module corresponding to Module-80 are related to DNA damage, DNA repair, DNA integration etc.

Conclusions

We used the deep belief network to process the gene expression data and search for transcriptomic modules. One complexity that exists when using deep belief network is the parameter setting, i.e., how to set the proper number of hidden layers and number of nodes in each hidden layer. As there exists no "gold standard", people usually test different settings manually in order to find a setting that achieves a good performance. In this work, we found that for a given data set, if you gradually increased the number of nodes in each hidden layer, the number of nodes that were actually activated in each hidden layer was bounded by a certain number. In this work, we used those bounds to set the number of nodes in each hidden layer, which resulted in a good performance in terms of finding transcriptomic modules that are biologically meaningful.

The genetic perturbation data obtained from the gene deletions is a valuable resource for studying signaling pathways. The basic idea is that if the deletion of a gene perturbs a signaling pathway, then the expression levels of genes regulated by the pathway will change significantly. By comparing the expression profiles, we could obtain relevant information to decide if the deletions of two individual genes perturb a common signal. However, as 1) there exist some random perturbations, or even just because of that cells may be in the different phases of cell cycle, and 2) a gene/protein, such as CDC42, in a pathway can take roles in other pathways [42]. Hence, besides the genes regulated by the pathway, some other genes can also be differentially expressed in the gene deletion experiments, which causes problems if we are comparing the genome-wide expression profiles. In this work, in order to greatly reduce the above problems, we first found transcriptomic modules such that genes in each module are highly likely to be regulated by a common pathway. We then only compared the expression profiles on genes in transcriptomic modules. Our results showed that utilizing the intrinsic relation between the perturbation of a pathway and the expression changes of genes regulated by the pathway is very helpful for studying the signaling systems.

There exist other data sets that used small molecules or drugs to perturb cell signaling systems and obtained the expression profile changes of cells, such as CMap data [43, 44] and LINCS data [45, 46]. As those expression data were basically obtained from single perturbation, our new computational model can also be applied to those data to find what small molecules or drugs perturb the same pathway. As a result, clinicians have the option to target other genes in a pathway if targeting one gene in this pathway does not work for a patient in targeted therapy.

Supporting information

S1 Fig. A global enrichment analysis of top 20 most enriched GO terms and their matched transcriptomic modules. In order to more easily view the result, we have taken negative log values of enrichment p-values and then normalize them.
(EPS)

S2 Fig. A global enrichment analysis of top 20 most enriched GO terms and their matched perturbation modules. In order to more easily view the result, we have taken negative log values of enrichment p-values and then normalize them.
(EPS)

S3 Fig. Comparing the GO enrichment analysis of perturbation modules obtained from using binarized fold-change, original fold-change, and normalized fold-change.
(EPS)

Acknowledgments

This work was supported by U.S. National Library of Medicine (Grant R00LM011673 to Dr. Songjian Lu) and National Human Genome Research Institute (Grant U54HG008540 to Xinghua Lu). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: Songjian Lu, Xinghua Lu.

Data curation: Songjian Lu.

Formal analysis: Xiaonan Fan.

Funding acquisition: Songjian Lu.

Investigation: Songjian Lu.

Methodology: Songjian Lu.

Project administration: Songjian Lu.

Software: Songjian Lu, Lujia Chen.

Validation: Xiaonan Fan.

Writing – original draft: Songjian Lu.

Writing – review & editing: Songjian Lu, Xinghua Lu.

References

1. Sebastian-Leon P, Vidal E, Minguez P, Conesa A, Tarazona S, Amadoz A, et al. Understanding disease mechanisms with models of signaling pathway activities. BMC systems biology. 2014; 8:121. https://doi.org/10.1186/s12918-014-0121-3 PMID: 25344409; PubMed Central PMCID: PMC4213475.
2. Whittaker S, Marais R, Zhu AX. The role of signaling pathways in the development and treatment of hepatocellular carcinoma. Oncogene. 2010; 29(36):4989–5005. https://doi.org/10.1038/onc.2010.236 PMID: 20639898.
3. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. Oncogene. 2007; 26 (22):3279–90. https://doi.org/10.1038/sj.onc.1210421 PMID: 17496922.
4. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. Cell. 2000; 102(1):109–26. PMID: 10929718.
5. Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, Margaritis T, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell. 2014; 157(3):740–52. https://doi.org/10.1016/j.cell.2014.02.054 PMID: 24766815.
6. Lenstra TL, Benschop JJ, Kim T, Schulze JM, Brabers NA, Margaritis T, et al. The specificity and topology of chromatin interaction pathways in yeast. Molecular cell. 2011; 42(4):536–49. https://doi.org/10.1016/j.molcel.2011.03.026 PMID: 21596317; PubMed Central PMCID: PMC4435841.
7. Liu Y, Zhao H. A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. BMC bioinformatics. 2004; 5:158. https://doi.org/10.1186/1471-2105-5-158 PMID: 15504238; PubMed Central PMCID: PMC526379.
8. Hu X, Wei H, Zheng H. Identification of perturbed signaling pathways from gene expression data using information divergence. Molecular bioSystems. 2017. https://doi.org/10.1039/c7mb00285h PMID: 28702621.

9. Steffen M, Petti A, Aach J, D'Haeseleer P, Church G. Automated modelling of signal transduction networks. BMC bioinformatics. 2002; 3:34. PMID: 12413400; PubMed Central PMCID: PMC137599.

10. Zhao J, Zhang S, Wu LY, Zhang XS. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics. 2012; 28(22):2940–7. https://doi.org/10.1093/bioinformatics/bts564 PMID: 22982574.

11. Huang R, Wallqvist A, Covell DG. Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. Genomics. 2006; 87(3):315–28. https://doi.org/10.1016/j.ygeno.2005.11.011 PMID: 16386875.

12. Brosch T, Tam R. Efficient training of convolutional deep belief networks in the frequency domain for application to high-resolution 2D and 3D images. Neural computation. 2015; 27(1):211–27. https://doi.org/10.1162/NECO_a_00682 PMID: 25380341.

13. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006; 313(5786):504–7. https://doi.org/10.1126/science.1127647 PMID: 16873662.

14. Huang GB, Lee H, Learned-Miller E, editors. Learning hierarchical representations for face verification with convolutional deep belief networks. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012; Washington, DC, USA.

15. Ch'ng SI, Seng KP, Ang LM. Block-based Deep Belief Networks for face recognition. International Journal of Biometrics. 2012; 4(2).

16. Herce HD, Deng W, Helma J, Leonhardt H, Cardoso MC. Visualization and targeted disruption of protein interactions in living cells. Nature communications. 2013; 4:2660. https://doi.org/10.1038/ncomms3660 PMID: 24154492; PubMed Central PMCID: PMC3826628.

17. Mayer BJ. Protein-protein interactions in signaling cascades. Methods in molecular biology. 2006; 332:79–99. https://doi.org/10.1385/1-59745-048-0:79 PMID: 16878686.

18. Voncken JW, Schweizer D, Aagaard L, Sattler L, Jantsch MF, van Lohuizen M. Chromatin-association of the Polycomb group protein BMI1 is cell cycle-regulated and correlates with its phosphorylation status. Journal of cell science. 1999; 112 (Pt 24):4627–39. PMID: 10574711.

19. McMillan JN, Tatchell K. The JNM1 gene in the yeast Saccharomyces cerevisiae is required for nuclear migration and spindle orientation during the mitotic cell cycle. The Journal of cell biology. 1994; 125 (1):143–58. PMID: 8138567; PubMed Central PMCID: PMC2120013.

20. Wang Y, Hu F, Elledge SJ. The Bfa1/Bub2 GAP complex comprises a universal checkpoint required to prevent mitotic exit. Current biology: CB. 2000; 10(21):1379–82. PMID: 11084339.

21. Vaisica JA, Baryshnikova A, Costanzo M, Boone C, Brown GW. Mms1 and Mms22 stabilize the replisome during replication stress. Molecular biology of the cell. 2011; 22(13):2396–408. https://doi.org/10.1091/mbc.E10-10-0848 PMID: 21593207; PubMed Central PMCID: PMC3128540.

22. Dovey CL, Russell P. Mms22 preserves genomic integrity during DNA replication in Schizosaccharomyces pombe. Genetics. 2007; 177(1):47–61. https://doi.org/10.1534/genetics.107.077255 PMID: 17660542; PubMed Central PMCID: PMC2013719.

23. Bi X, Barkley LR, Slater DM, Tateishi S, Yamaizumi M, Ohmori H, et al. Rad18 regulates DNA polymerase kappa and is required for recovery from S-phase checkpoint-mediated arrest. Molecular and cellular biology. 2006; 26(9):3527–40. https://doi.org/10.1128/MCB.26.9.3527-3540.2006 PMID: 16611994; PubMed Central PMCID: PMC1447421.

24. Zhu XD, Kuster B, Mann M, Petrini JH, de Lange T. Cell-cycle-regulated association of RAD50/MRE11/NBS1 with TRF2 and human telomeres. Nature genetics. 2000; 25(3):347–52. https://doi.org/10.1038/77139 PMID: 10888888.

25. Gatei M, Jakob B, Chen P, Kijas AW, Becherel OJ, Gueven N, et al. ATM protein-dependent phosphorylation of Rad50 protein regulates DNA repair and cell cycle control. The Journal of biological chemistry. 2011; 286(36):31542–56. https://doi.org/10.1074/jbc.M111.258152 PMID: 21757780; PubMed Central PMCID: PMC3173097.

26. Lisby M, Rothstein R, Mortensen UH. Rad52 forms DNA repair and recombination centers during S phase. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98 (15):8276–82. https://doi.org/10.1073/pnas.121006298 PMID: 11459964; PubMed Central PMCID: PMC37432.

27. Xu C, Wang Y, Wang L, Wang Q, Du LQ, Fan S, et al. Accumulation and Phosphorylation of RecQ-Mediated Genome Instability Protein 1 (RMI1) at Serine 284 and Serine 292 during Mitosis. International journal of molecular sciences. 2015; 16(11):26395–405. https://doi.org/10.3390/ijms161125965 PMID: 26556339; PubMed Central PMCID: PMC4661824.

28. Balogun FO, Truman AW, Kron SJ. DNA resection proteins Sgs1 and Exo1 are required for G1 checkpoint activation in budding yeast. DNA repair. 2013; 12(9):751–60. https://doi.org/10.1016/j.dnarep.2013.06.003 PMID: 23835406; PubMed Central PMCID: PMC3769955.

29. Mankouri HW, Hickson ID. Top3 processes recombination intermediates and modulates checkpoint activity after DNA damage. Molecular biology of the cell. 2006; 17(10):4473–83. https://doi.org/10.1091/mbc.E06-06-0516 PMID: 16899506; PubMed Central PMCID: PMC1635375.

30. Sharp JA, Rizki G, Kaufman PD. Regulation of histone deposition proteins Asf1/Hir1 by multiple DNA damage checkpoint kinases in Saccharomyces cerevisiae. Genetics. 2005; 171(3):885–99. https://doi.org/10.1534/genetics.105.044719 PMID: 16020781; PubMed Central PMCID: PMC1456847.

31. Hu F, Alcasabas AA, Elledge SJ. Asf1 links Rad53 to control of chromatin assembly. Genes & development. 2001; 15(9):1061–6. https://doi.org/10.1101/gad.873201 PMID: 11331602; PubMed Central PMCID: PMC312686.

32. Clausing E, Mayer A, Chanarat S, Muller B, Germann SM, Cramer P, et al. The transcription elongation factor Bur1-Bur2 interacts with replication protein A and maintains genome stability during replication stress. The Journal of biological chemistry. 2010; 285(53):41665–74. https://doi.org/10.1074/jbc.M110.193292 PMID: 21075850; PubMed Central PMCID: PMC3009894.

33. Fumasoni M, Zwicky K, Vanoli F, Lopes M, Branzei D. Error-free DNA damage tolerance and sister chromatid proximity during DNA replication rely on the Polalpha/Primase/Ctf4 Complex. Molecular cell. 2015; 57(5):812–23. https://doi.org/10.1016/j.molcel.2014.12.038 PMID: 25661486; PubMed Central PMCID: PMC4352764.

34. Crabbe L, Thomas A, Pantesco V, De Vos J, Pasero P, Lengronne A. Analysis of replication profiles reveals key role of RFC-Ctf18 in yeast replication stress response. Nature structural & molecular biology. 2010; 17(11):1391–7. https://doi.org/10.1038/nsmb.1932 PMID: 20972444.

35. Xu H, Boone C, Klein HL. Mrc1 is required for sister chromatid cohesion to aid in recombination repair of spontaneous damage. Molecular and cellular biology. 2004; 24(16):7082–90. https://doi.org/10.1128/MCB.24.16.7082-7090.2004 PMID: 15282308; PubMed Central PMCID: PMC479732.

36. Karras GI, Jentsch S. The RAD6 DNA damage tolerance pathway operates uncoupled from the replication fork and is functional beyond S phase. Cell. 2010; 141(2):255–67. https://doi.org/10.1016/j.cell.2010.02.028 PMID: 20403322.

37. Hedglin M, Benkovic SJ. Regulation of Rad6/Rad18 Activity During DNA Damage Tolerance. Annual review of biophysics. 2015; 44:207–28. https://doi.org/10.1146/annurev-biophys-060414-033841 PMID: 26098514.

38. Chahwan C, Nakamura TM, Sivakumar S, Russell P, Rhind N. The fission yeast Rad32 (Mre11)-Rad50-Nbs1 complex is required for the S-phase DNA damage checkpoint. Molecular and cellular biology. 2003; 23(18):6564–73. https://doi.org/10.1128/MCB.23.18.6564-6573.2003 PMID: 12944482; PubMed Central PMCID: PMC193710.

39. Roset R, Inagaki A, Hohl M, Brenet F, Lafrance-Vanasse J, Lange J, et al. The Rad50 hook domain regulates DNA damage signaling and tumorigenesis. Genes & development. 2014; 28(5):451–62. https://doi.org/10.1101/gad.236745.113 PMID: 24532689; PubMed Central PMCID: PMC3950343.

40. Sidorova JM, Breeden LL. Rad53-dependent phosphorylation of Swi6 and down-regulation of CLN1 and CLN2 transcription occur in response to DNA damage in Saccharomyces cerevisiae. Genes & development. 1997; 11(22):3032–45. PMID: 9367985; PubMed Central PMCID: PMC316703.

41. Mohanty S, Town T, Yagi T, Scheidig C, Kwan KY, Allore HG, et al. Defective p53 engagement after the induction of DNA damage in cells deficient in topoisomerase 3beta. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(13):5063–8. https://doi.org/10.1073/pnas.0801235105 PMID: 18367668; PubMed Central PMCID: PMC2278186.

42. Oehlen LJ, Cross FR. The role of Cdc42 in signal transduction and mating of the budding yeast Saccharomyces cerevisiae. The Journal of biological chemistry. 1998; 273(15):8556–9. PMID: 9535827.

43. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313 (5795):1929–35. https://doi.org/10.1126/science.1132939 PMID: 17008526.

44. Lamb J. The Connectivity Map: a new tool for biomedical research. Nature reviews Cancer. 2007; 7 (1):54–60. https://doi.org/10.1038/nrc2044 PMID: 17186018.

45. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics. 2016; 32(15):2338–45. https://doi.org/10.1093/bioinformatics/btw168 PMID: 27153606; PubMed Central PMCID: PMC4965635.

46. Vempati UD, Chung C, Mader C, Koleti A, Datar N, Vidovic D, et al. Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the Library of Integrated Network-based Cellular Signatures (LINCS). Journal of biomolecular screening. 2014; 19(5):803–16. https://doi.org/10.1177/1087057114522514 PMID: 24518066.