Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Novel Method for the Prediction of Drug-Drug Interactions Based on Gene Expression Profiles

Y-H. TAUGCHI[1], and TURKI TURKI[2]
[1]Department of Physics, Chuo University, Tokyo 112-8551, Japan (e-mail: tag@granular.com)
[2]King Abdulaziz University, Department of Computer Science, Jeddah, 21589, Saudi Arabia (e-mail: tturki@kau.edu.sa)

Corresponding author: Y-h. Taguchi (e-mail: tag@granular.com).

The study was supported by KAKENHI, 19H05270, 20H04848, and 20K12067. This project was also funded by the Deanship of Scientific Research (DSR) at King Ab-dulaziz University, Jeddah, under grant no. KEP-8-611-38. The authors thank DSR for technical and financial support.

ABSTRACT The accurate prediction of new interactions between drugs is important for avoiding unknown (mild or severe) adverse reactions to drug combinations. The development of effective *in silico* methods for evaluating drug interactions based on gene expression data requires an understanding of how various drugs alter gene expression. Current computational methods for the prediction of drug-drug interactions (DDIs) utilize data for known DDIs to predict unknown interactions. However, these methods are limited in the absence of known predictive DDIs. To improve DDIs' interpretation, a recent study has demonstrated strong non-linear (i.e., dose-dependent) effects of DDIs. In this study, we present a new unsupervised learning approach involving tensor decomposition (TD)-based unsupervised feature extraction (FE) in 3D. We utilize our approach to reanalyze available gene expression profiles for *Saccharomyces cerevisiae*. We found that non-linearity is possible, even for single drugs. Thus, non-linear dose-dependence cannot always be attributed to DDIs. Our analysis provides a basis for the design of effective methods for evaluating DDIs.

INDEX TERMS Bioinformatics, Drug-drug interaction, Feature extraction, Gene expression, Tensor decomposition, Unsupervised learning,

I. INTRODUCTION

ALTHOUGH *in silico* methods are thought to be effective strategies for improving the long, expensive process of drug discovery, *in silico* drug discovery is, at best, still under development [1]–[3]. In addition to the two main approaches for drug discovery, i.e., ligand-based drug discovery [4]–[6] and structure-based drug discovery [7]–[9], interest in gene expression profile-based drug discovery [10] has recently increased. For this process, it is important to understand how drug treatments alter gene expression profiles. However, this is a complex issue owing to the huge number of gene expression alterations resulting from each treatment. The alterations are often non-linear, with non-monotonic dose-dependent effects. This non-linearity often prevents the selection of effective drugs, since it is difficult to determine if expression levels of individual genes are up- or downregulated by particular drug treatments. In drug discovery, analyses of drug-drug interactions (DDIs) are aimed at the prevention or reduction of possible reactions caused by therapeutic drug combinations [11]–[17]. Several machine-learning approaches have been proposed to predict interactions between drugs accurately. For example, Yan et al. [18] proposed a learning approach called DDIGIP, which utilizes a regularized least square classifier coupled with a Gaussian interaction profile (GIP) kernel on known DDI profiles to predict new DDIs; the performance of this approach was supported by 5-fold (and 10-fold) experimental cross-validation. Rohani et al. [19] proposed a learning approach utilizing a neural network in which concatenated pairs of drugs are used as inputs according to a calculated integrated similarity matrix to predict unknown interactions. Experimental results have demonstrated that the proposed approach performs better than other baselines. Other learning approaches have also been proposed to predict new DDIs [11], [20]–[26]. However, the above-mentioned methods are not capable of predicting unknown interactions if data for known DDIs are not available. Hence, Lukačišin and Bollenbach [27] evaluated how DDIs affect gene expression profiles in a combinatorial man-

ner; they found that DDIs can exhibit convex relationships with gene expression profiles. Our main contributions are summarized as follows. (1) We provide a method for the reliable interpretation of the effects of interactions between drugs on gene expression data; in particular, we propose a new unsupervised method involving tensor decomposition (TD)-based unsupervised feature extraction (FE) [28] and apply this approach to datasets used in [27]. (2) We demonstrate that our TD-based unsupervised FE can replicate the findings of Lukačišin and Bollenbach [27] based on a principal component analysis (PCA) [29]. (3) Using the newly proposed TD-based unsupervised FE method, we show that convex dose-dependence can appear in single-drug treatments. Thus, our analysis improves our general understanding of DDIs in [27], especially when considering multi-drug effects. (4) As our analysis provides detailed insight into interactions between drugs in the context of gene expression [30], it has practical implications for improving performance when designing computational methods to predict interactions between drugs accurately.

II. MATERIALS AND METHODS

Fig. 1 shows a diagram of the analyses performed in this study.

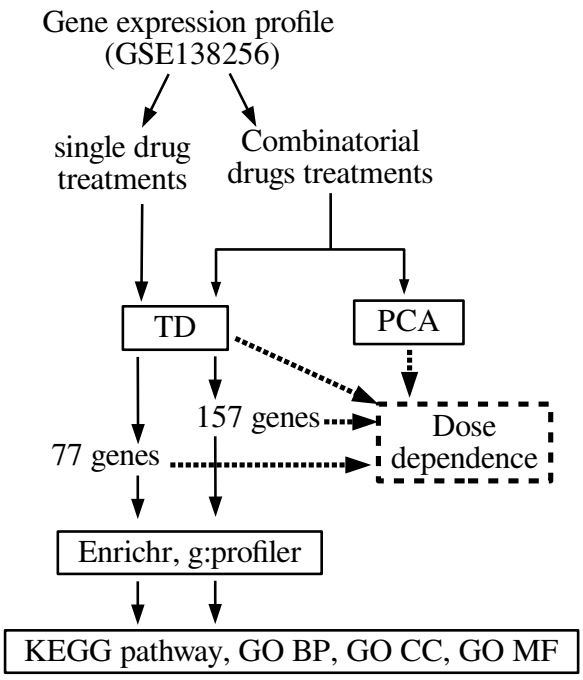


FIGURE 1. Flowchart of analyses performed in this study.

A. 2.1 GENE EXPRESSION PROFILES

Gene expression profiles were downloaded from Gene Expression Omnibus (GEO) [31] with GEO ID GSE138256. The processed file named "GSE138256_GeneExpression.csv.gz" was used. The file "GSE138256_SampleConditionsAndOrdering.csv.g was also downloaded for sample annotations. These dataset sets were composed of gene expression profiles of *Saccharomyces cerevisiae* treated with individual drugs or pairs of the following four drugs: myriocin, cycloheximide, LiCl, and rapamycin. When *S. cerevisiae* was treated with pairs of drugs, and the combinatorial dose was carefully tuned to ensure the same growth rate, to the greatest extent possible.

TABLE 1. Number of doses tested for drug combinations.

	Myriocin	Rapamycin	LiCl
Cycloheximide	22	20	19
LiCl	27	18	
Rapamycin	16		

TABLE 2. Number of doses tested for individual drugs. Numbers in parentheses indicate unique doses.

drug	Number of samples
Myriocin	25 (14)
Cycloheximide	23 (11)
LiCl	28 (14)
Rapamycin	30 (14)

B. PCA

PCA was applied to individual pairs of drugs. For the ith gene expression level and jth dose, $x_{ij} \in \mathbb{R}^{N\times M}$, where N is total number of genes (i.e., 6717) and M is total number of combinatorial doses for each pair of drugs (Table 1). x_{ij} is normalized as $\sum_i x_{ij} = 0$ and $\sum_i x_{ij}^2 = N$. PCA was applied to x_{ij} such that PC loadings and PC scores were attributed to samples and genes, respectively. Lowess smoothing was applied to PC loadings to reduce noise signals using the lowess command implemented in R [32].

C. TD-BASED UNSUPERVISED FE

TD-based unsupervised FE was applied to gene expression profiles. Gene expression profiles were formatted as a tensor, $x_{ijk} \in \mathbb{R}^{N\times 16\times 6}$, representing the expression of the ith gene and jth combinatorial dose of the kth pair of drugs. Since the number of combinatorial doses varied among pairs, the minimum number of combinatorial doses, 16, was employed. When more combinatorial doses were tested for specific pairs of drugs, some measurements were discarded, attempting to maintain equal intervals between doses. x_{ijk} was normalized as $\sum_i x_{ijk} = 0$ and $\sum_i x_{ijk}^2 = N$. Higher order singular value decomposition (HOSVD) [28] was applied to x_{ijk} to obtain the following:

$$x_{ijk} = \sum_{\ell_1\ell_2\ell_3} G(\ell_1\ell_2\ell_3) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 i} \tag{1}$$

where $G(\ell_1\ell_2\ell_3) \in \mathbb{R}^{N\times 16\times 6}$ is a core tensor, and $u_{\ell_1 j} \in \mathbb{R}^{16\times 16}$, $u_{\ell_2 k} \in \mathbb{R}^{6\times 6}$, and $u_{\ell_3 i} \in \mathbb{R}^{N\times N}$ are the singular value vectors defined as the column vectors of orthogonal matrices. $u_{\ell_1 j}$ is attributed to the jth dose, $u_{\ell_2 k}$ is attributed to the kth pair of drugs, and $u_{\ell_3 i}$ is attributed to the ith gene.

Lowess smoothing was also applied to singular value vectors to reduce noise using the lowess command implemented in R [32]. To select $u_{\ell_3 i}$ for gene selection in subsequent analyses, it was first necessary to determine which $u_{\ell_1 j}$ and $u_{\ell_2 k}$ are biologically meaningful. After identifying such ℓ_1 and ℓ_2, it is necessary to identify the ℓ_3 associated with $G(\ell_1 \ell_2 \ell_3)$ having the largest absolute values given fixed ℓ_1 and ℓ_2. With the selected ℓ_3, P-values, P_i, were obtained for the ith gene as follows:

$$P_i = P_{\chi^2}\left[> \sum \left(\frac{u_{\ell_3 i}}{\sigma_{\ell_3}}\right)^2\right] \tag{2}$$

where $P_{\chi^2}[> \ x]$ is the cumulative probability of the χ^2 distribution with an argument larger than x. The summation is taken over only selected ℓ_3s to compute P-values. P_i was corrected with the BH criterion [28] and genes with $P < 0.01$ were selected.

Gene expression levels in response to a single dose (Table 2) were also formatted as a three-mode tensor, $x_{ijk} \in \mathbb{R}^{N \times 14 \times 4}$, which represents the ith gene expression level for the jth dose of the kth drug. Since the number of unique doses was 14, excluding cycloheximide, the total number of doses for cycloheximide was also set to 14, and two replicates were included for three doses. The same procedure employed for the analysis of combinatorial drug treatments was repeated, and genes were selected.

D. ENRICHMENT ANALYSIS

The gene symbols of selected genes were uploaded to YeastEnrichr, a yeast version of Enricher [33], prepared for humans, as well as to g:profiler [34]

III. RESULTS

We first applied PCA to gene expression levels, x_{ij}, attributed to individual pairs of drug treatments to attempt to reproduce the previous observations [27]. In the previous study [27], the first PC loading takes constant values, independent of dose, while the second and the third PC loadings exhibit linear and convex dose-dependence, regardless of pairs of drugs. In our analysis, the first PC also took constant values, irrespective of the drug combination (not shown). However, the second and third PC loadings behaved slightly differently (Fig. 2). For the combination of cycloheximide and LiCl, although the second and the third PC loadings behaved as expected, the fourth PC loading also showed concave or convex dose-dependence. Since the fourth PC loadings were not discussed in the previous study [27], it is possible that the observation was observed in the previous research but was not reported. Nevertheless, for the combination of LiCl and rapamycin, the second PC loadings did not have linear dependence but instead showed stepwise dependence, which was not reported in the previous study [27]. Additionally, for the remaining four combinatorial cases, the second and third PCs did not always have linear and concave or convex dose-dependence, respectively.

It is possible that the disagreement between the present study (in which the third PC did not always have linear and concave or convex dependence on the dose, respectively) and the previous study [27] could be explained by insufficient pre-processing of gene expression profiles. To evaluate this possibility, we applied HOSVD to the tensor, x_{ijk}, generated from combinatorial drug treatments (for a summary of this analysis, see Fig. 3). It was evident that u_{1j} takes constant values independent of dose density, u_{2j} has a linear dependence on dose density, and u_{3j} has concave or convex dependence on dose density, as observed in the original study [27] (Fig. 4). This suggests the superiority of TD-based unsupervised FE to identify essential features, regardless of pre-processing.

One might notice that u_{4j} is also concave or convex and u_{5j} and u_{6j} have more complex shapes (S-letter shaped). To see if these shapes are artifacts or reflect individual gene expression profiles, we focused on genes whose expression levels are likely coincident with these concave and convex shapes. Since we noticed that u_{1k} had constant values over six combinatorial treatments, we searched for $G(\ell_1 \ 1 \ \ell_3)$ with the largest contribution to $\ell_1 = 3, 4$ and relatively small contributions to $\ell_1 = 1, 2$, which are associated with constant or linear dependence (Fig. 5A). It is obvious that $G(\ell_1 \ 1 \ 1)$ had the largest contribution to $\ell_1 = 1$, i.e., a constant (or dose density-independent) profile, while $G(\ell_1 \ 1 \ 2)$ and $G(\ell_1 \ 1 \ 3)$ had the largest contribution to $\ell_1 = 2$, i.e., linearly dependent on dose density. Thus, to identify $u_{\ell_3 i}$ associated with profiles other than constant or linear profiles, we employed $4 \leq \ell_3 \leq 6$ for gene selection. Based on P-values and correction as described in the Materials and Methods, we selected 157 genes (Table 3).

TABLE 3. List of 157 genes selected by TD-based unsupervised FE toward combinatorial drugs treatments. These genes are associated with concave or convex dose-dependence, since they are expected to be associated with u_{3j} and u_{4j} (Fig. 4).

BDH1 GCV3 CDC19 YAL037C-B SSA1 ADE1 YBL005W-B TIP1 HSP26 YBR116C TKL2 TEF2 DUR1 GLK1 HIS4 AGP1 PGK1 YCR013C HSP30 RPL35A RPL41A RPL41B TPI1 FMP16 YDR154C CPR1 HSP42 YDR210C-D HSP78 YDR261C-D HXT7 HXT6 HXT3 RPS17B EMI2 YDR524W-C YRF1-1 DLD3 GLC3 GCN4 TIR1 RGI1 YER067C-A YER138C YRF1-2 ACT1 HSP12 GSY1 RPL29 YFR032C-B YFR052C-A HXK1 ADE5 YGL102C OLE1 LEU1 YGR027W-A YGR038C-B NQM1 CTT1 TPO2 TDH3 ADE3 ENO1 BGL2 YRF1-3 YHL050C YHR052W-A CUP1-1 YHR054W-A CUP1-2 HXT4 RPL42B ENO2 YHR219W RPL39 BUD19 RPS21B YJL133C-A TDH1 TDH2 OPI3 SOD1 BAT2 GPM1 CWP2 CWP1 FBA1 UGP1 YLL066C UBI4 HSP104 SSA2 YLR035C-A PDC1 SHM2 RPL22A BUD28 AHP1 CCW12 YLR154W-A YLR154W-B YLR154W-F YLR154C-G YLR157C-B YLR162W RRT15 RPS31 CBF5 NOP56 YLR198C YLR227W-B YEF3 RPL38 ADE13 YRF1-4 YRF1-5 YML133C DAK1 YML039W TSA1 GLO1 YMR045C YMR046W-A PGM2 ADE17 ALD3 SIP18 HSC82 YRF1-6 DBP2 LEU4 POR1 YNL054W-B RPL25 ADH1 RPS30B WTM1 RPS12 GDH1 FIT3 YRF1-8 YRF1-7 HSP82 SSE1 YPR002C-A GLN1 RPL43A OPI11 TEF1 RPS23B YPR137C-B ASN1 YPR158C-D YPR158C-C GPH1 YPR204W

To see if the 157 genes selected in this analysis were associated with concave, convex, or the more complicated S-shaped pattern, we plotted Lowess-smoothed expression profiles of two representative genes, BDH1 and SSA1, as shown

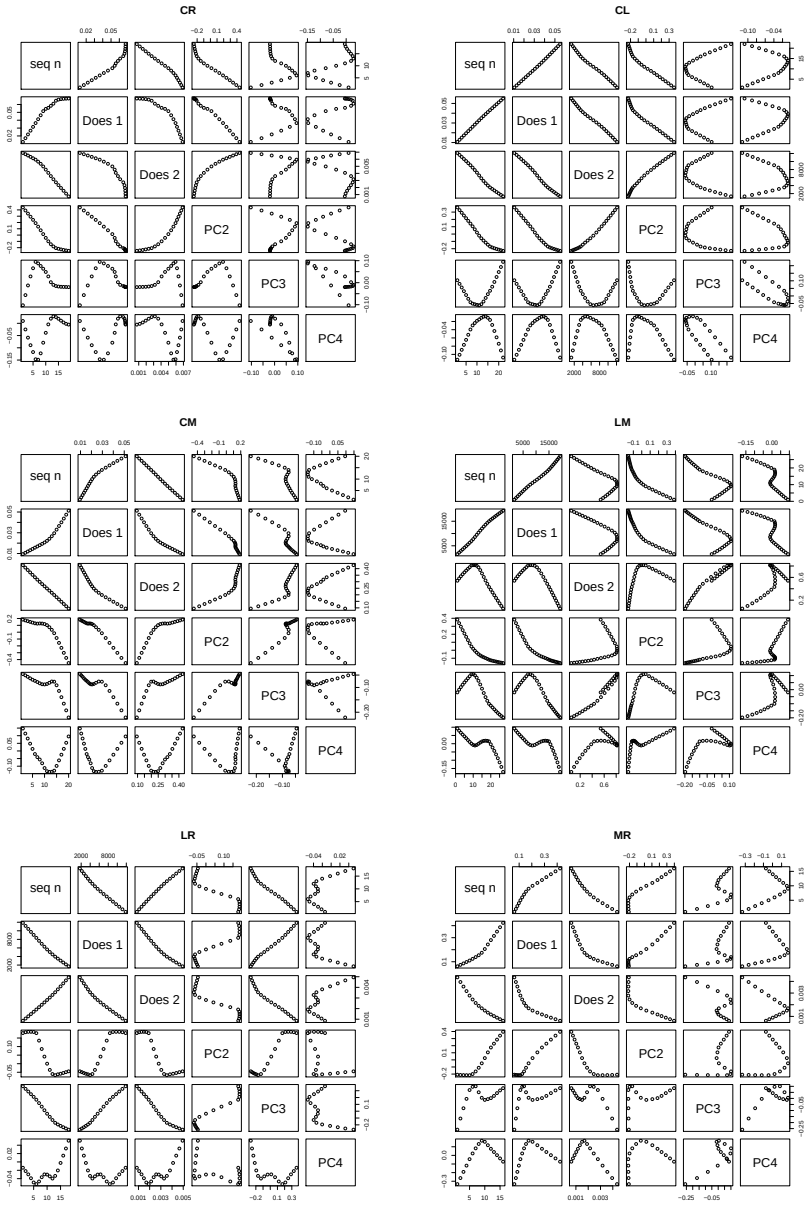


FIGURE 2. Scatter plots of j, dose densities of the first and second drug, and the second to fourth PC loadings. All values are Lowess-smoothed. Sequential numbers are ordered along iso-growth rate lines identified in the previous study [27] as much as possible. Two letters above each panel show the combinations of drugs: M: Myriocin C: Cycloheximide, L: LiCl, R: Rapamycin

in Fig. 6 (note that gene expression profiles of other genes are available as supplementary materials). Gene expression profiles have distinct dose-dependence for drug combinations, although concave, convex, and S-shaped profiles were observed. Thus, the profiles shown in Fig. 4 were not artifacts but reflected the expression patterns of individual genes. TD-based unsupervised FE not only generated singular value vectors that represent constant, linear, concave, or convex dependence on dose density but also characterizes more complicated (S-letter shaped) profiles for individual genes. Thus, it is an advantageous strategy for analyzing gene expression profiles obtained under distinct conditions in an integrated manner.

Next, we validated the selected genes by evaluating their biological functions. We uploaded 157 genes to YeastEnrichr and found enrichment for numerous biological functions. In particularly, we detected 23 significant biological terms in the KEGG pathway analysis (see 10 top-ranked terms in Table 4), 91 terms in the GO Biological Process (BP) category (see 10 top-ranked terms in Table 5), 22 terms in the GO Cellular Component (CC) category (see 10 top-ranked terms in Table 6), and 35 terms in the GO Molecular Function (MF) category (see 10 top-ranked terms in Table 7). Thus, the selected genes had key biological functions. To confirm

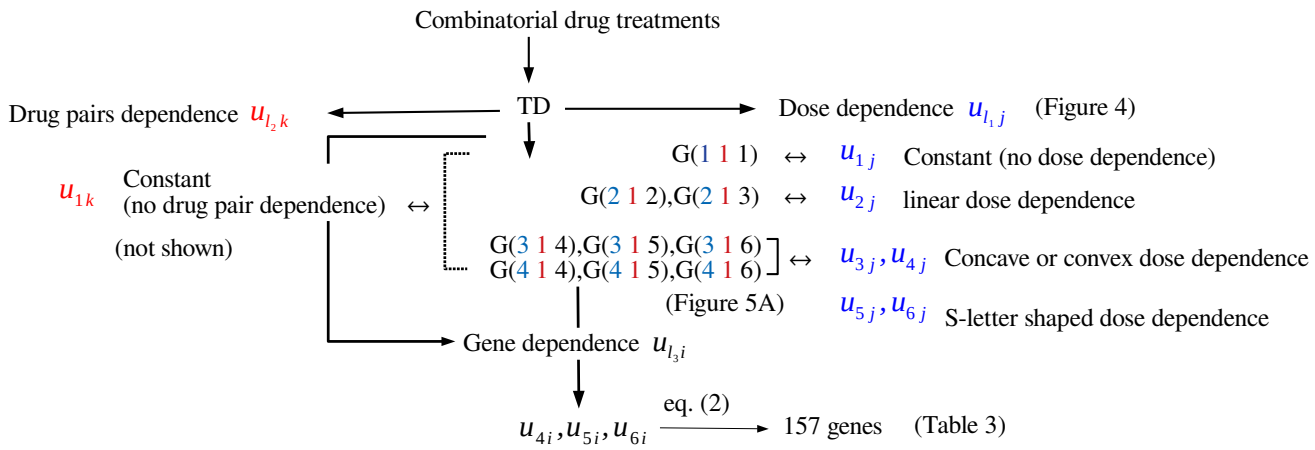


FIGURE 3. Summary of TD-based unsupervised FE applied to gene expression with combinatorial drug treatments

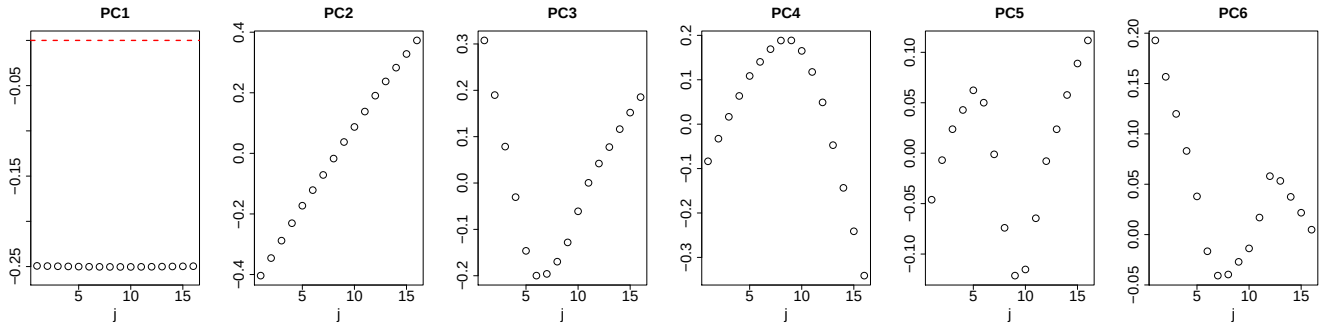


FIGURE 4. Lowess-smoothed $u_{\ell_1 j}, 1 \leq \ell_1 \leq 6$ for combinatorial drug treatments.

the observed enrichment, we also analyzed the genes using g:profiler. Although we obtained fewer significantly enriched terms, there were 219 biological terms, including KEGG pathways and GO BP, MF, and CC terms (lists of individual biological terms obtained using YeastEnrichr and g:profiler are available as supplementary materials). Thus, the biological significance of the selected genes is not database-dependent, supporting the robustness and reliability of the analysis.

IV. DISCUSSION AND CONCLUSION

In this study, we partially reproduced the original observations [27] by PCA; however, TD-based unsupervised FE allowed us to obtain the same results more robustly and reliably. Based on our findings, the expression levels of some genes exhibit non-linear dependence on the dose density. However, non-linear dependence on the dose density was also observed for treatment with single drugs (see Additional file 2 [35]). Thus, it is not clear whether the concave or convex dependence on the dose can be explained by DDIs. To further evaluate the ability of individual drugs to result in non-linear dose-dependence, we applied the newly developed TD-based unsupervised FE to the alternative tensor, x_{ijk}, generated from gene expression profiles of *S. cerevisiae* treated with single drugs (see Materials and Methods and Fig. 7 for the summary of this analysis). Fig. 8 shows the Lowess-smoothed $u_{\ell_1 j}, 1 \leq \ell_1 \leq 6$ for single-drug treatments. Contrary to our expectation, non-linearity was substantially greater than that shown in Fig. 4 based on combinatorial treatments. Linear dependence was minimal, and an S-letter shaped pattern was observed prior to concave or convex patterns. To determine if the strong non-linearity is associated with individual gene expression profiles, we selected genes associated with singular value vectors that exhibit non-linearity, as shown in Fig. 6. We initially noticed that u_{1k} has constant values over four single-drug treatments, as in the case of combinatorial drug treatments (not shown). Thus we need to find $G(\ell_1\ 1\ \ell_3)$ with the largest contribution to $3 \leq \ell_1 \leq 6$ and relatively small contributions to $\ell_1 = 1, 2$, indicating constant or linear dependence (Fig. 8).

Observed patterns (Fig. 5B) exhibited greater non-linearity than those shown in Fig. 5A for combinatorial treatments. When drugs were treated in combinatorial manner, $G(1\ 1\ 1)$ has the largest absolute values among $G(1\ 1\ \ell_3)$; this means

TABLE 4. Ten top-ranked KEGG 2019 terms for 157 genes selected by TD-based unsupervised FE when combinatorial drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
Glycolysis / Gluconeogenesis	17/54	2.47×10^{-23}	1.14×10^{-21}
Starch and sucrose metabolism	9/39	1.56×10^{-11}	2.39×10^{-10}
Ribosome	16/170	3.81×10^{-13}	8.76×10^{-12}
Fructose and mannose metabolism	6/21	1.05×10^{-8}	1.20×10^{-7}
Methane metabolism	6/25	3.33×10^{-8}	3.06×10^{-7}
Galactose metabolism	5/22	6.61×10^{-7}	5.07×10^{-6}
Amino sugar and nucleotide sugar metabolism	5/30	3.40×10^{-6}	2.23×10^{-5}
Protein processing in endoplasmic reticulum	7/88	5.99×10^{-6}	3.44×10^{-5}
Longevity regulating pathway	5/36	8.66×10^{-6}	4.43×10^{-5}
Valine, leucine and isoleucine biosynthesis	3/12	9.91×10^{-5}	3.26×10^{-4}

TABLE 5. Ten top-ranked GO biological process (BP) 2018 terms for 157 genes selected by TD-based unsupervised FE when combinatorial drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
glycolytic process (GO:0006096)	13/20	1.92×10^{-23}	3.06×10^{-21}
ATP generation from ADP (GO:0006757)	13/20	1.92×10^{-23}	3.06×10^{-21}
nicotinamide nucleotide metabolic process (GO:0046496)	13/24	6.00×10^{-22}	6.38×10^{-20}
pyruvate metabolic process (GO:0006090)	14/33	1.35×10^{-21}	1.07×10^{-19}
carbohydrate catabolic process (GO:0016052)	13/28	8.77×10^{-21}	5.59×10^{-19}
glucose metabolic process (GO:0006006)	13/32	7.92×10^{-20}	4.21×10^{-18}
gluconeogenesis (GO:0006094)	9/16	9.80×10^{-16}	3.91×10^{-14}
hexose biosynthetic process (GO:0019319)	9/16	9.80×10^{-16}	3.91×10^{-14}
cytoplasmic translation (GO:0002181)	16/162	1.79×10^{-13}	6.35×10^{-12}
translation (GO:0006412)	20/297	2.22×10^{-13}	7.07×10^{-12}

TABLE 6. Ten top-ranked GO cellular component (CC) 2018 terms for 157 genes selected by TD-based unsupervised FE when combinatorial drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
cytosolic part (GO:0044445)	20/204	1.64×10^{-16}	5.64×10^{-15}
cytosol (GO:0005829)	32/676	1.82×10^{-16}	5.64×10^{-15}
retrotransposon nucleocapsid (GO:0000943)	15/91	4.40×10^{-16}	9.10×10^{-15}
nucleus (GO:0005634)	44/1599	7.80×10^{-14}	1.21×10^{-12}
cytosolic ribosome (GO:0022626)	17/185	1.00×10^{-13}	1.25×10^{-12}
mitochondrion (GO:0005739)	33/1063	8.36×10^{-12}	8.64×10^{-11}
fungal-type cell wall (GO:0009277)	12/132	5.56×10^{-10}	4.93×10^{-9}
cytosolic large ribosomal subunit (GO:0022625)	10/101	6.94×10^{-9}	5.38×10^{-8}
large ribosomal subunit (GO:0015934)	10/104	9.24×10^{-9}	6.36×10^{-8}
cytosolic small ribosomal subunit (GO:0022627)	6/71	2.00×10^{-5}	1.24×10^{-4}

TABLE 7. Ten top-ranked GO molecular function (MF) 2018 terms for 157 genes selected by TD-based unsupervised FE when combinatorial drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
helicase activity (GO:0004386)	12/39	1.16×10^{-16}	1.30×10^{-14}
RNA-directed DNA polymerase activity (GO:0003964)	12/48	1.95×10^{-15}	1.09×10^{-13}
DNA-directed DNA polymerase activity (GO:0003887)	12/59	2.91×10^{-14}	1.09×10^{-12}
DNA polymerase activity (GO:0034061)	12/61	4.47×10^{-14}	1.25×10^{-12}
nuclease activity (GO:0004518)	12/64	8.27×10^{-14}	1.85×10^{-12}
ribonuclease activity (GO:0004540)	12/66	1.22×10^{-13}	2.28×10^{-12}
RNA binding (GO:0003723)	24/477	4.39×10^{-13}	7.02×10^{-12}
DNA helicase activity (GO:0003678)	8/35	2.37×10^{-10}	3.32×10^{-9}
purine ribonucleoside triphosphate binding (GO:0035639)	7/78	2.66×10^{-6}	3.31×10^{-5}
nucleoside-triphosphatase activity (GO:0017111)	10/195	3.33×10^{-6}	3.74×10^{-5}

that constant profiles are associated with the first singular value vector, $u_{\ell i}$. $G(2\ 1\ 3)$, $G(2\ 1\ 5)$, and $G(2\ 1\ 6)$ had the largest absolute values among $G(2\ 1\ \ell_3)$, indicating that linear profiles are associated with the second singular value vector, u_{2i}, as well as the third singular value vector, u_{3i}. Nevertheless, in Fig. 5B, although $G(1\ 1\ 1)$ had the largest absolute values among $G(1\ 1\ \ell_3)$s, $G(2\ 1\ \ell_3), 2 \leq \ell_3 \leq 6$, had substantial contributions, indicating that there is no clear separation between genes whose expression profiles are associated with dose-dependence represented by u_{2i}, which are most likely linear profiles, and those with dose-dependence represented by u_{3i} to u_{6i}, likely representing non-linear profiles, i.e., concave, convex, and S-letter shaped profiles. Thus, to select genes with strong non-linear dependence on

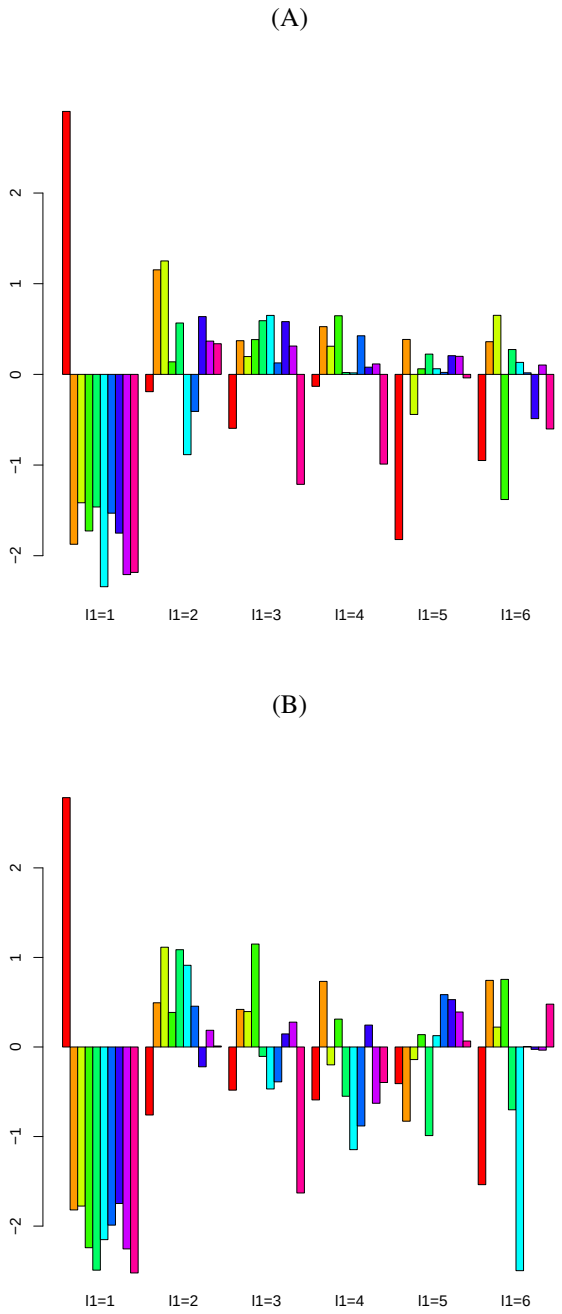


FIGURE 5. Common logarithmic absolute values of $G(\ell_1\ 1\ \ell_3)$ for combinatorial (A) or single (B) drug treatments. For each ℓ_1, $G(\ell?_1\ 1\ \ell_3)$ values are aligned from left to right in increasing order of ℓ_3. The same colors correspond to the same ℓ_3.

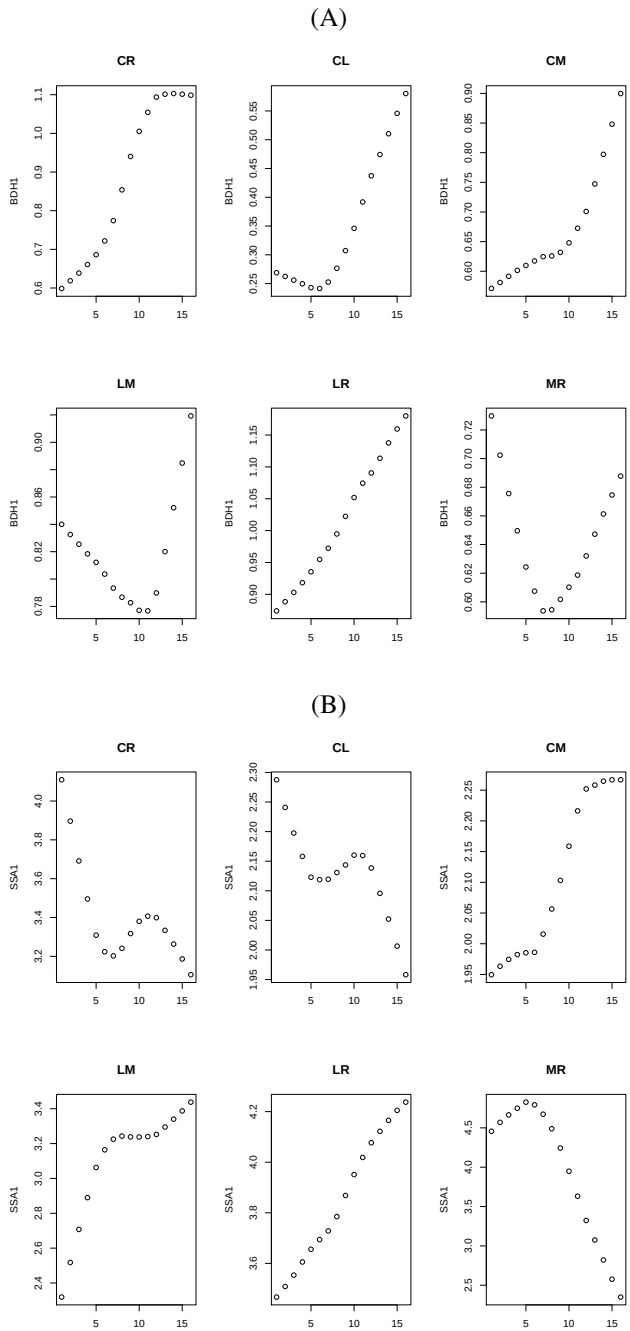


FIGURE 6. Lowess-smoothed gene expression profiles for BDH1 (A) and SSA1 (B). Two letters above each panel show the combinations of drugs: M: Myriocin C: Cycloheximide, L: LiCl, R: Rapamycin

dose, we selected $\ell_3 = 4$, since $G(3\ 1\ 4)$ had the largest absolute values among $G(3\ 1\ \ell_3)$ and ultimately identified 77 significant genes (Table 8).

Fig. 9 shows Lowess-smoothed gene expression profiles of two representative genes whose expression levels are also shown in Fig. 6 with respect to combinatorial drug treatments (expression profiles of other genes are available as supplementary materials). Non-linearity of dose-dependence is not clearly reduced. Accordingly, the strong non-linearity of dose-dependence observed for combinatorial drug treatments may not reflect DDIs but rather the non-linearity on the dose-dependence of the expression of individual genes (as shown in Fig. 10, showing an extensive overlap of selected genes for single and combinatorial drug treatments). In conclusion, in our comparison of gene expression profiles between single

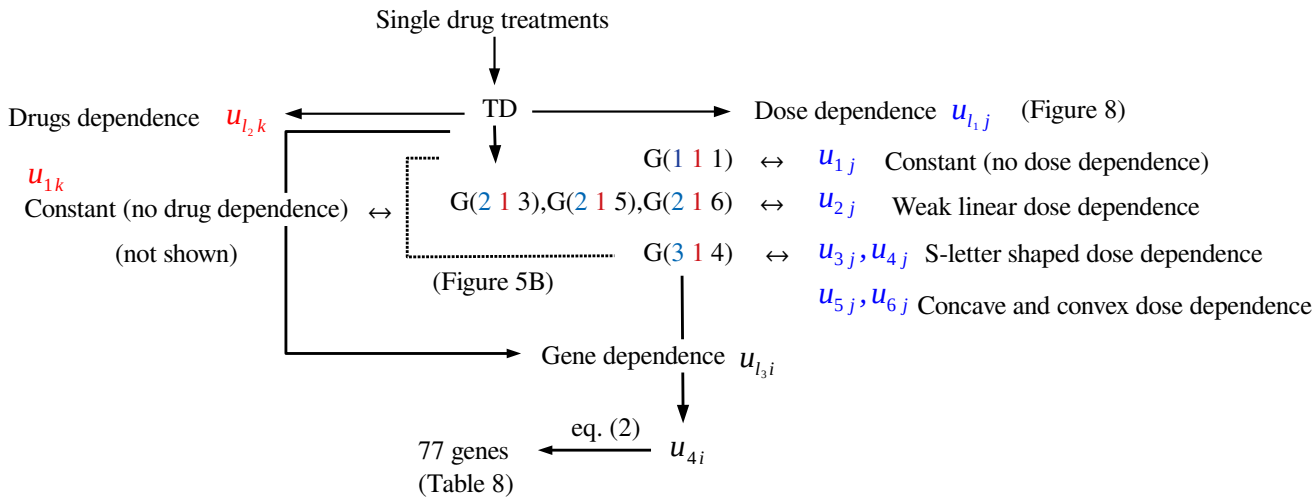


FIGURE 7. Summary of TD-based unsupervised FE applied to gene expression with single drug treatments

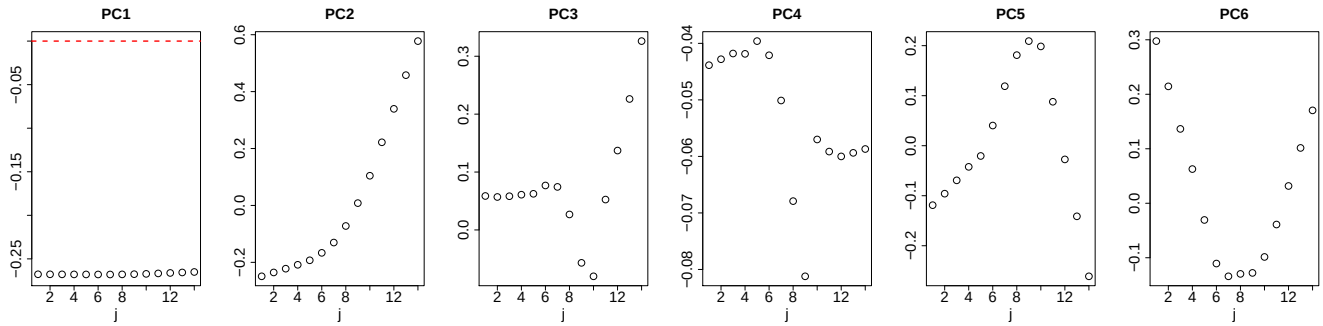


FIGURE 8. Lowess-smoothed $u_{\ell_1 j}, 1 \leq \ell_1 \leq 6$ for single drug treatments.

TABLE 8. List of 77 genes selected by TD-based unsupervised FE toward single drugs treatments. These genes are associated with non-linear dose-dependence, since they are expected to be associated with u_{3j} (Fig. 8).

BDH1 SSA1 YAR009C RPS8A YBL005W-B YBR012W-B HSP26 RPS6B GLK1 PGK1 YCR013C YCR018C-A TPI1 HSP42 YDR261C-D YDR316W-B HXT7 HXT6 EMI2 OM45 CYC7 GLC3 RGI1 YER067C-A YER138C YER160C HSP12 YFR052C-A HXK1 PNC1 STF2 YGR038C-B CTT1 YGR161C-D TDH3 ENO1 YHR052W-A CUP1-1 YHR054W-A CUP1-2 RTC3 HXT4 YHR214C-B TDH1 TDH2 SOD1 YKL153W GPM1 FBA1 UGP1 HSP104 YLR035C-A CCW12 YLR157C-B TFS1 YLR227W-B YEF3 TMA10 YML045W YML039W YMR045C PGM2 ALD3 YMR173W-A YNL284C-B RPS3 YNL054W-B YNR034W-A ADH1 RPS12 YPL257W-B HSP82 RPS6A YPR137C-B YPR158W-B YPR158C-D GPH1

and combinatorial drug treatments, we did not obtain clear evidence that the strong non-linearity between gene expression levels and dose can be directly attributed to DDIs.

We further evaluated the biological significance of the 77 genes selected for treatment with single drugs. We identified a number of significant (adjusted P-values less than 0.05) KEGG pathways (Table 9) and GO terms in the BP (Table 10), CC (Table 11), and MF categories (Table 12). Thus, the selected genes were biologically relevant.

We confirmed the observed patterns of enrichment using g:profiler. In this analysis, we detected fewer significantly enriched terms overall but still observed enrichment for various KEGG pathways and GO terms. Thus, the biological significance of the selected genes did not depend on the database, and the analyses were robust and reliable (lists of individual biological terms obtained using YeastEnrichr and g:profiler are available as supplementary materials).

To analyze and interpret the effects of drug interactions on gene expression, we propose a new unsupervised method, a TD-based unsupervised FE in 3D, and applied it to gene expression profiles of *S. cerevisiae* treated with single or combinatorial drugs. Because strong non-linear dependence was observed for both treatments (separate and combined), our analysis demonstrates that these effects are unlikely to reflect DDIs.

One might wonder why we did not compare our methods with other methods. First of all, we compare our methods with PCA that was employed in [27]. Thus, it is miss-interpretation that we did not compare ours with no other

TABLE 9. Ten top-ranked KEGG 2019 terms for 77 genes selected by TD-based unsupervised FE when single drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
Glycolysis / Gluconeogenesis	14/54	1.30×10^{-22}	3.76×10^{-21}
Starch and sucrose metabolism	7/39	1.32×10^{-10}	1.92×10^{-9}
Fructose and mannose metabolism	5/21	1.44×10^{-8}	1.34×10^{-7}
Galactose metabolism	5/22	1.85×10^{-8}	1.34×10^{-7}
Amino sugar and nucleotide sugar metabolism	5/30	9.80×10^{-8}	5.68×10^{-7}
Longevity regulating pathway	5/36	2.55×10^{-7}	1.23×10^{-6}
Methane metabolism	3/25	1.19×10^{-4}	4.92×10^{-4}
Protein processing in endoplasmic reticulum	4/88	3.71×10^{-4}	1.34×10^{-3}
Ribosome	5/170	5.04×10^{-4}	1.63×10^{-3}
Tyrosine metabolism	2/14	1.29×10^{-3}	3.75×10^{-3}

TABLE 10. Ten top-ranked GO BP 2018 terms for 77 genes selected by TD-based unsupervised FE when single drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
glycolytic process (GO:0006096)	11/20	2.13×10^{-22}	1.69×10^{-20}
ATP generation from ADP (GO:0006757)	11/20	2.13×10^{-22}	1.69×10^{-20}
nicotinamide nucleotide metabolic process (GO:0046496)	11/24	3.13×10^{-21}	1.65×10^{-19}
carbohydrate catabolic process (GO:0016052)	11/28	2.66×10^{-20}	1.05×10^{-18}
glucose metabolic process (GO:0006006)	11/32	1.58×10^{-19}	5.00×10^{-18}
pyruvate metabolic process (GO:0006090)	11/33	2.36×10^{-19}	6.23×10^{-18}
gluconeogenesis (GO:0006094)	8/16	4.17×10^{-16}	8.23×10^{-15}
hexose biosynthetic process (GO:0019319)	8/16	4.17×10^{-16}	8.23×10^{-15}
glucose import (GO:0046323)	5/33	1.62×10^{-7}	2.84×10^{-6}
glucose transport (GO:0015758)	5/34	1.89×10^{-7}	2.99×10^{-6}

TABLE 11. Ten top-ranked GO CC 2018 terms for 77 genes selected by TD-based unsupervised FE when single drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
retrotransposon nucleocapsid (GO:0000943)	22/91	1.62×10^{-34}	6.30×10^{-33}
nucleus (GO:0005634)	34/1599	9.70×10^{-18}	1.89×10^{-16}
cytosol (GO:0005829)	17/676	5.94×10^{-10}	7.73×10^{-9}
cytosolic part (GO:0044445)	8/204	1.18×10^{-6}	1.15×10^{-5}
cytosolic small ribosomal subunit (GO:0022627)	5/71	7.92×10^{-6}	6.18×10^{-5}
mitochondrion (GO:0005739)	15/1063	1.10×10^{-5}	7.12×10^{-5}
small ribosomal subunit (GO:0015935)	5/79	1.34×10^{-5}	7.45×10^{-5}
cytosolic ribosome (GO:0022626)	6/185	7.94×10^{-5}	3.87×10^{-4}
fungal-type cell wall (GO:0009277)	5/132	1.57×10^{-4}	6.79×10^{-4}
chaperonin-containing T-complex (GO:0005832)	2/12	9.42×10^{-4}	3.67×10^{-3}

TABLE 12. Ten top-ranked GO MF 2018 terms for 77 genes selected by TD-based unsupervised FE when single drug treatments were employed.

Term	Overlap	P-value	Adjusted P-value
RNA-directed DNA polymerase activity (GO:0003964)	21/48	2.04×10^{-39}	1.04×10^{-37}
DNA-directed DNA polymerase activity (GO:0003887)	21/59	4.61×10^{-37}	1.18×10^{-35}
DNA polymerase activity (GO:0034061)	21/61	1.08×10^{-36}	1.83×10^{-35}
nuclease activity (GO:0004518)	21/64	3.60×10^{-36}	4.60×10^{-35}
ribonuclease activity (GO:0004540)	21/66	7.77×10^{-36}	7.92×10^{-35}
RNA binding (GO:0003723)	25/477	5.41×10^{-22}	4.60×10^{-21}
glucokinase activity (GO:0004340)	3/6	1.09×10^{-6}	5.55×10^{-6}
hexokinase activity (GO:0004396)	3/6	1.09×10^{-6}	5.55×10^{-6}
mannokinase activity (GO:0019158)	3/6	1.09×10^{-6}	5.55×10^{-6}
fructokinase activity (GO:0008865)	3/6	1.09×10^{-6}	5.55×10^{-6}

methods. Second, since our approach is an unsupervised method, to our knowledge, there are no other methods to achieve that we have had in this study; i. e., investigate what the representative gene expression profiles without assuming anything and compare the representative gene expression profiles with individual gene expression profiles. Thus, we could not find anything to be compared with our methods other than PCA.

One might also mind if the present study is novel enough to be published since I have even a whole book on this TD-based unsupervised FE. Nevertheless, TD-based unsupervised FE is too general to deny individuals because a dual application is nothing but an application of this method. TD-based unsupervised FE could be used to various topics ranging from biomarker identification, diseases causing genes identification, and drug discovery. Since Application to DDI is nothing but a new application of TD-based unsupervised FE, we do not think that the present study does not have

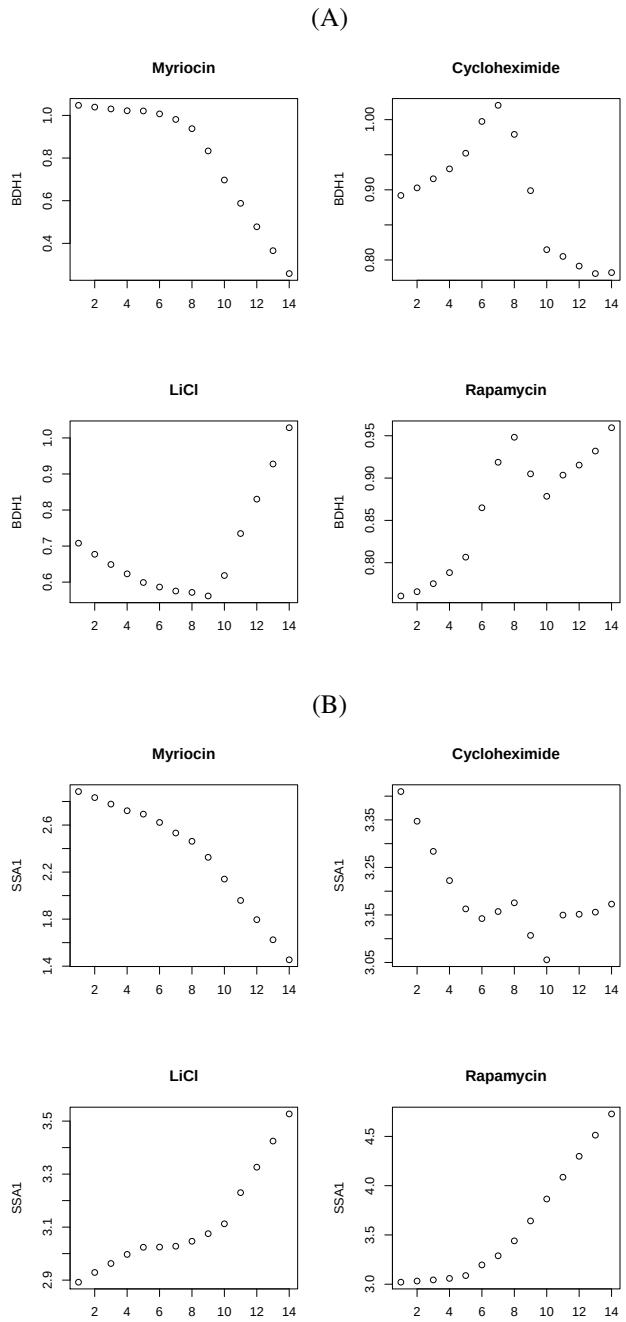


FIGURE 9. Lowess-smoothed gene expression profiles for BDH1 (A) and SSA1 (B). Two letters above each panel show the combinations of drugs: M: Myriocin C: Cycloheximide, L: LiCl, R: Rapamycin

enough novelty to be published.

The reason why we employed the yeast as a target organism is because we do have data for other organisms. When there are data sets, we can apply our methods to these data sets.

At the moment, we do not have any limitations to our methods. We believe that we can apply our technique to any kind of gene expression profiles obtained by the combinatorial treatments of multiple drugs. We are waiting for data set which are publicly available.

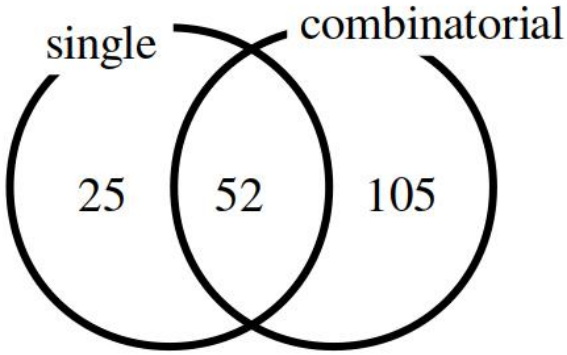


FIGURE 10. Venn diagram of genes selected for combinatorial and single drug treatments.

REFERENCES

[1] Ahmet Sureyya Rifaioglu, Heval Atas, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Briefings in Bioinformatics, 20(5):1878–1912, 07 2018.

[2] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6):463–477, apr 2019.

[3] Sayada Reemsha Kazmi, Ren Jun, Myeong-Sang Yu, Chanjin Jung, and Dokyun Na. In silico approaches and tools for the prediction of drug metabolism and fate: A review. Computers in Biology and Medicine, 106:54 – 64, 2019.

[4] Magdalena Bacilieri and Stefano Moro. Ligand-based drug design methodologies in drug discovery process: An overview. Current Drug Discovery Technologies, 3(3):155–165, 2006.

[5] Sourav Pal, Vinay Kumar, Biswajit Kundu, Debomita Bhattacharya, Nagothy Preethy, Mamindla Prashanth Reddy, and Arindam Talukdar. Ligand-based pharmacophore modeling, virtual screening and molecular docking studies for discovery of potential topoisomerase i inhibitors. Computational and Structural Biotechnology Journal, 17:291 – 310, 2019.

[6] Matthew C. Robinson, Robert C. Glen, and Alpha A. Lee. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. Journal of Computer-Aided Molecular Design, 34(7):717–730, jan 2020.

[7] Maria Batool, Bilal Ahmad, and Sangdun Choi. A structure-based drug discovery paradigm. International Journal of Molecular Sciences, 20(11), 2019.

[8] Y. h. Taguchi. Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets. Scientific Reports, 7(1), oct 2017.

[9] Yoonji Lee, Raudah Lazim, Stephani Joy Y Macalino, and Sun Choi. Importance of protein dynamics in the structure-based drug discovery of class a g protein-coupled receptors (gpcrs). Current Opinion in Structural Biology, 55:147 – 153, 2019.

[10] Murty V. Chengalvala, Vargheese M. Chennathukuzhi, Daniel S. Johnston, Panayiotis E. Stevis, and Gregory S. Kopf. Gene expression profiling and its practice in drug development. Current Genomics, 8(4):262–270, 2007.

[11] Remzi Celebi, Huseyin Uyar, Erkan Yasar, Ozgur Gumus, Oguz Dikenelli, and Michel Dumontier. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. BMC Bioinformatics, 20(1), dec 2019.

[12] Xiaohui Yao, Tiffany Tsang, Qing Sun, Sara Quinney, Pengyue Zhang, Xia Ning, Lang Li, and Li Shen. Mining and visualizing high-order

directional drug interaction effects using the FAERS database. BMC Medical Informatics and Decision Making, 20(S2), mar 2020.

[13] Jian-Yu Shi, Xue-Qun Shang, Ke Gao, Shao-Wu Zhang, and Siu-Ming Yiu. An integrated local classification model of predicting drug-drug interactions via dempster-shafer theory of evidence. Scientific Reports, 8(1), aug 2018.

[14] Aleksandar Poleksic and Lei Xie. Database of adverse events associated with drugs and drug combinations. Scientific Reports, 9(1), dec 2019.

[15] Juanhong Zhang, Yuemei Sun, Rong Wang, and Junmin Zhang. Gut microbiota-mediated drug-drug interaction between amoxicillin and aspirin. Scientific Reports, 9(1), nov 2019.

[16] Jacob A. Langness and Gregory T. Everson. Drug–drug interactions in HCV treatment — the good, the bad and the ugly. Nature Reviews Gastroenterology & Hepatology, 13(4):194–195, feb 2016.

[17] Yosef Masoudi-Sobhanzadeh, Yadollah Omidi, Massoud Amanlou, and Ali Masoudi-Nejad. Drugr+: A comprehensive relational database for drug repurposing, combination therapy, and replacement therapy. Computers in Biology and Medicine, 109:254 – 262, 2019.

[18] Cheng Yan, Guihua Duan, Yi Pan, Fang-Xiang Wu, and Jianxin Wang. DDIGIP: predicting drug-drug interactions based on gaussian interaction profile kernels. BMC Bioinformatics, 20(S15), dec 2019.

[19] Narjes Rohani and Changiz Eslahchi. Drug-drug interaction predicting by neural network using integrated similarity. Scientific Reports, 9(1), sep 2019.

[20] Adeeb Noor, Abdullah Assiri, Serkan Ayvaz, Connor Clark, and Michel Dumontier. Drug-drug interaction discovery and demystification using Semantic Web technologies. Journal of the American Medical Informatics Association, 24(3):556–564, 12 2016.

[21] Dalong Song, Yao Chen, Qian Min, Qingrong Sun, Kai Ye, Changjiang Zhou, Shengyue Yuan, Zhaolin Sun, and Jun Liao. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. Journal of Clinical Pharmacy and Therapeutics, 44(2):268–275, 2019.

[22] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. Proceedings of the National Academy of Sciences, 115(18):E4304–E4311, 2018.

[23] Wen Zhang, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian, and Xiaohong Li. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinformatics, 18(1), jan 2017.

[24] Feixiong Cheng and Zhongming Zhao. Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. Journal of the American Medical Informatics Association, 21(e2):e278–e286, 03 2014.

[25] Yi Zheng, Hui Peng, Xiaocai Zhang, Zhixun Zhao, Xiaoying Gao, and Jinyan Li. DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions. BMC Bioinformatics, 20(S19), dec 2019.

[26] Geonhee Lee, Chihyun Park, and Jaegyoon Ahn. Novel deep learning model for more accurate prediction of drug-drug interaction effects. BMC Bioinformatics, 20(1), aug 2019.

[27] Martin Lukačišin and Tobias Bollenbach. Emergent gene expression responses to drug combinations predict higher-order drug interactions. Cell Systems, 9(5):423 – 433.e3, 2019.

[28] Unsupervised feature extracion applied to bioinformatics: A PCA based and TD based approach. Springer international, 2019.

[29] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202, apr 2016.

[30] Leslie Z. Benet, Christine M. Bowman, Megan L. Koleske, Capria L. Rinaldi, and Jasleen K. Sodhi. Understanding drug–drug interaction and pharmacogenomic changes in pharmacokinetics for metabolized drugs. Journal of Pharmacokinetics and Pharmacodynamics, 46(2):155–163, mar 2019.

[31] Emily Clough and Tanya Barrett. The gene expression omnibus database. In Methods in Molecular Biology, pages 93–110. Springer New York, 2016.

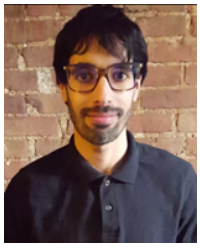
[32] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[33] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research, 44(W1):W90–W97, 05 2016.

[34] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Research, 47(W1):W191–W198, 05 2019.

[35] Y h. Taguchi. Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data. BMC Bioinformatics, 19(S13), feb 2019.

Y-H. TAGUCHI received a B.S. degree in physics from the Tokyo Institute of Technology and a Ph.D. degree in physics from the Tokyo Institute of Technology. He is currently a full professor with the Department of Physics, Chuo University, Japan. His works have been published in leading journals such as Physical Review Letters, Bioinformatics, and Scientific Reports. His research interests include bioinformatics, machine-learning, and non-linear physics. He is also an editorial board member of Frontiers in Genetics:RNA, PloS ONE, BMC Medical Genomics, Medicine (Lippincott Williams & Wilkins journal), BMC Research Notes, non-coding RNA (MDPI), and IPSJ Transaction on Bioinformatics.



TURKI TURKI received a B.S. in computer science from King Abdulaziz University, an M.S. in computer science from NYU.POLY, and a Ph.D. in computer science from the New Jersey Institute of Technology. He is currently an assistant professor with the Department of Computer Science, King Abdulaziz University, Saudi Arabia. His research interests include artificial intelligence, machine learning, deep learning, data mining, data science, big data analytics, and bioinformatics. His research has been accepted and published in journals such as Frontiers in Genetics, BMC Genomics, BMC Systems Biology, Expert Systems with Applications, Computers in Biology and Medicine, and Current Pharmaceutical Design. He was awarded several distinction awards from the Deanship of Scientific Research at King Abdulaziz University. He is supported by King Abdulaziz University and is currently working on several biomedicine related projects. Dr. Turki has served on the program committees of several international conferences. Additionally, he is an editorial board member of Sustainable Computing: Informatics and Systems and Computers in Biology and Medicine.