Neurological disorder drug discovery from gene expression with tensor decomposition

Y-h. Taguchi† and Turki Turki⋆

†Department of Physics, Chuo University, Tokyo 112-8551, Japan
⋆King Abdulaziz University, Department of Computer Science, Jeddah 21589, Saudi Arabia

July 15, 2019

Abstract

Background: Identifying effective candidate drug compounds in patients with neurological disorders based on gene expression data is of great importance to the neurology field. By identifying effective candidate drugs to a given neurological disorder, neurologists would (1) reduce the time searching for effective treatments; and (2) gain additional useful information that leads to a better treatment outcome. Although there are many strategies to screen drug candidate in pre-clinical stage, it is not easy to check if candidate drug compounds can be also effective to human.
Objective: We tried to propose a strategy to screen genes whose expression is altered in model animal experiments to be compared with gene expressed differentically with drug treatment to human cell lines.
Methods: Recently proposed tensor decomposition (TD) based unsupervised feature extraction (FE) is applied to single cell (sc) RNA-seq experiments of Alzheimer's disease model animal mouse brain.
Results: Four hundreds and one genes are screened as those differentially expressed during Aβ accumulation as age progresses. These genes are significantly overlapped with those expressed differentially with the known drug treatments for three independent data sets: LINCS, DrugMatrix and GEO.
Conclusion: Our strategy, application of TD based unsupervised FE, is useful one to screen drug candidate compounds using scRNA-seq data set.

keywords: Amyloid, Alzheimer Disease, Gene Expression, Single-Cell Analysis, Drug Discovery, Cell Line

1 Introduction

Drug discovery for neurological disorder has never been successful in spite of massive efforts spent [1]. One possible reason is because we generally do not

have suitable model animals for human neurological disorder [2]. Although a huge number of compounds are screened using model animals, only a few of them passed the human level screening. In this sense, it is required to screen candidate compounds using information retrieved from human at the earliest stage. One possible strategy to do this is the usage of human cell lines; Nevertheless, it is also not easy to perform, since generating cell line from human neurological disorder patients is not easy. In contrast to the cancer cell lines, which can be easily generated by immortalizing tumor cells, neuronal cells are hardly converted to cell lines, since mature neurons do not undergo cell division [3]. Therefore, it is difficult to test if candidate drugs work for human during pre-clinical stages.

In order to overcome this difficulty, we proposed an alternative strategy; comparing disease gene expression with that of compound treated animals and/or human cell lines. Generally, compound screening is based upon phenotype; i.e., evaluation of compounds efficiency is tested based upon if drug treatment can produce symptomatic improvement. Nevertheless, since it has been recently found that various neurological disorders share gene expression [4], focusing on gene expression profiles might be more reasonable. Following this strategy, we considered gene expression profiles (single cell RNA-seq) of mouse brain during amyloid β accumulation. As being aged, some set of gene expression progresses and significantly overlaps with genes that express differential expression caused by various compounds treatment. Since top ranked (i.e., with the most overlaps) detected compounds turn out to be tested previously toward Alzheimer disease (AD) treatment, lower ranked compounds also might be promising candidate compounds for AD.

2 Materials and Methods

2.1 Single cell RNA-seq

Single cell (sc) RNA-seq used in this study was downloaded from gene expression omnibus (GEO) using GEO ID GSE127892. It is composed of two genotypes (APP_NL-F-G and C57Bl/6), two tissues (Cortex and Hippocampus), four ages (3, 6, 12, and 21 weeks), two sex (male and female) and four 96 well plates. For each of combined combinations, four 96 well plates, each of wells includes one cell, were tested. Among those wells tested, wells with insufficient gene expression were discarded. As a result, among 2 (genotype) $\times$ 2 (tissues) $\times$ 4 (ages) $\times$ 2 (sex) $\times$ 4 (plates) $\times$ 96 (wells) = 12288 cells measured, scRNA-seq for only 10801 cells were provided.

2.2 Tensor decomposition based unsupervised feature extraction

We applied recently proposed tensor decomposition (TD) based unsupervised feature extraction (FE) [5–14] to scRNA-seq. A tensor $x_{j_1 j_2 j_3 j_4 j_5 j_6 i} \in \mathbb{R}^{96\times 2\times 2\times 4\times 2\times 4\times 29341}$

that represents gene expression of ith gene of j_1th cell (well) at j_2th genotyoe ($j_2 = 1$:APP_NL-F-G and $j_2 = 2$: C57Bl/6), j_3th tissue ($j_3 = 1$:Cortex and $j_3 = 2$:Hippocampus), j_4th age ($j_4 = 1$: three weeks, $j_4 = 2$: six weeks, $j_4 = 3$: twelve weeks, and $j_4 = 4$: twenty one weeks), j_5th sex ($j_5 = 1$:female and $j_5 = 2$:male) and j_6th plate.

$x_{j_1 j_2 j_3 j_4 j_5 j_6 i}$ is standardized such that $\sum_{i=1}^{29341} x_{j_1 j_2 j_3 j_4 j_5 j_6 i} = 0$ and $\sum_{i=1}^{29341} x^2_{j_1 j_2 j_3 j_4 j_5 j_6 i} = 29341$. HOSVD [5] was applied to $x_{j_1 j_2 j_3 j_4 j_5 j_6 i}$ such that

$$x_{j_1 j_2 j_3 j_4 j_5 j_6 i} = \sum_{\ell_1=1}^{96} \sum_{\ell_2=1}^{2} \sum_{\ell_3=1}^{2} \sum_{\ell_4=1}^{4} \sum_{\ell_5=1}^{2} \sum_{\ell_6=1}^{4} \sum_{\ell_7=1}^{29341} G(\ell_1, \ell_2, \ell_3, \ell_4, \ell_5, \ell_6, \ell_7) u_{\ell_1 j_1} u_{\ell_2 j_2} u_{\ell_3 j_3} u_{\ell_4 j_4} u_{\ell_5 j_5} u_{\ell_6 j_6} u_{\ell_7 i} \tag{1}$$

where $G(\ell_1, \ell_2, \ell_3, \ell_4, \ell_5, \ell_6, \ell_7) \in \mathbb{R}^{96\times2\times2\times4\times2\times4\times29341}$ is core tensor, $u_{\ell_1 j_1} \in \mathbb{R}^{96\times96}$, $u_{\ell_2 j_2} \in \mathbb{R}^{2\times2}$, $u_{\ell_3 j_3} \in \mathbb{R}^{2\times2}$, $u_{\ell_4 j_4} \in \mathbb{R}^{4\times4}$, $u_{\ell_5 j_5} \in \mathbb{R}^{2\times2}$, $u_{\ell_6 j_6} \in \mathbb{R}^{4\times4}$ and $u_{\ell_6 i} \in \mathbb{R}^{29341\times29341}$ are singular value matrices that are orthogonal matrices. In order to save time to compute, only $1 \leq \ell_1, \ell_7 \leq 10$ were computed.

After investigation of $u_{\ell_4 j_4}$, $u_{2 j_4}$ represent monotonic dependence upon age while $\ell_1, \ell_2, \ell_3, \ell_5, \ell_6 = 1$ represent independence of cells, genotype, tissue, sex and plate. Since $G(1, 1, 1, 2, 1, 1, 2)$ has the largest absolute vales among $G(1, 1, 1, 2, 1, 1, \ell_7)$, $u_{2,i}$ is employed to compute P-values attributed to ith gene as

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

where $P_{\chi^2}[> x]$ is the cumulative probability of χ^2 distribution when the argument is larger than x and σ is the standard deviation.

P-values are corrcted by Benjamini and Hochberg criterion [15] and genes associated with corrected P-values less than 0.01 are selected for downstream analysis.

2.3 Enrichment analysis

Four hundreds and one genes selected by TD based unsupervised FE were uploaded to Enrichr [16] for enrichment analysis. Full list of enrichment analysis as well as list of 401 genes are accessible at
https://amp.pharm.mssm.edu/Enrichr3/enrich?dataset=5bbbe5602715daf9787895cd16829707

List of 401 genes and three enrichment analyses used in this study, "LINCS L1000 Chem Pert up", "DrugMatrx" and "Drug Perturbations from GEO up" are also available as supplementary material.

3 Results

When selecting genes using TD based unsupervised FE, we first need to specify what kind of properties of gene expression we consider. In this study, we require the followings.

1. Gene expression should be independent of cells within the same 96 wells plate.
2. Gene expression should be independent of genotype.
3. Gene expression should be independent of tissues.
4. Gene expression should have monotonic dependence upon age.
5. Gene expression should be independent of sex.
6. Gene expression should be independent of each of four 96 wells plates under the same conditions.

In other words, we try to select genes with the most robust monotonic age dependence as much as possible. The reason of this motivation is as follows. In the paper where data set analyzed here was investigated originally, Frigerio et al. [17] found that age is the primary factor of the microglia response to accumulation of Aβ plaques. We found that singular value vectors with $\ell_1 = \ell_2 = \ell_3 = \ell_5 = \ell_6 = 1$ represent independence of cells, genotypes, tissues, sex and plates (Figure 1 (A), (B), (C), (E), (F)). On the other hand, u_{2j_4} represents monotonic dependence upon ages, $1 \leq j_4 \leq 4$ (Figure 1 (D)).

Next, we need to find the $G(1, 1, 1, 2, 1, 1, \ell_7)$ with the largest absolute value in order to identify singular value vector, $u_{\ell_7 i}$, attributed to genes. Then we found that $G(1, 1, 1, 2, 1, 1, 2)$ has the largest absolute value. Therefore, we decided to use u_{2i} for attributing P-values to genes as shown in eq. (2). Finally, 401 genes are identified as being associated with adjusted P-values less than 0.01 (The list of genes is available as supplementary material).

These 401 genes are uploaded to Enrichr to identify the compounds, with which genes expressing differential expression of cell lines treated are maximally overlapped with these 401 genes. As for "LINCS L1000 Chem Pert up" category (Table 1, full list is available as supplementary material), the top ranked compound is alvocidib, which was previously tested for AD [18]; there are also 65 experiments (see supplementary material) of cell lines treated with alvocidib and associated with adjusted P-value less than 0.05. The second top ranked compound is AZD-8055, which was also previously tested for AD [19]; there are also 6 experiments (see supplementary material) of cell lines treated with AZD-8055 and associated with adjusted P-value less than 0.05.

One might wonder if this is an accidental agreement which is specific to LINCS data set. In order to confirm that it is not an accidental agreement, we also see DrugMatrix category (Table 2, full list is available as supplementary material). The top, fifth and tenth ranked compound is cyclosporin-A, which was also previously tested for AD [20];there are also 57 experiments (see supplementary material) of cell lines treated with cyclosporin-A and associated with adjusted P-value less than 0.05. Finally, we tested "Drug Perturbations from GEO up" category in Enrichr (Table 3, full list is available as supplementary material). The top ranked compounds is imatinib, which was also previously tested for AD [21];there are also 18 experiments (see supplementary material) of

cell lines treated with imatinib and associated with adjusted P-value less than 0.05.

These findings suggest that our strategy is effective to find compounds that can be used for AD treatment. Thus, it might be worthwhile investigating lower ranked compounds in Tables 1, 2 and 3 as candidate compounds for AD, even if they were not known drugs for AD.

4 Discussion

First of all, since these cell lines in Table 1 are originated in human, our strategy can provide us the opportunity to check if proposed candidate drugs screened with model animals are also effective in human.

It is also remarkable that we do not need gene expression of all genes, but only a subset of genes (please remember that LINCS project measures only gene expression of less than one thousand genes) in order to predict candidate drugs with high accuracy. This might reduce the amount of money to screen numerous number of compounds.

Our method is also applicable to scRNA-seq in order to screen drug compounds candidate from scRNA-seq. To our knowledge, there are very limited number of studies that relate scRNA-seq to drug design [22,23], since scRNA-seq usually lacks cell labeling which is useful to screen differentially expressed genes. In this study, we simply make use of ages, which is not always directly related to diseases. In spite of that, drug we listed was correct, i.e., they are known drugs to some extent. Therefore, our strategy is also useful to add an alternative one along this direction, i.e., making use of scRNA-seq for drug design.

Thus, our strategy, TD based unsupervised FE, might be promising methodology to screen drug candidate compounds.

5 Conclusion

In this paper, we applied TD based unsupervised FE to scRNA-seq taken from mouse brain with Aβ accumulation. We have compared selected 401 genes with differentially expressed genes in cell lines and model animals treated with various compounds. As a result, as for three independent data sets, LINCS, DrugMatrix and GEO, top ranked compounds are reported to be tested as AD treatment. This suggests the effectiveness of our strategy and lower ranked compounds should be tested as promising drug compounds candidates. To our knowledge, this is the first successful one that can be applied to scRNA-seq in order to identify drug compounds candidate.

6 Acknowledgement

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. KEP-8-611-38. The authors,

181 therefore, acknowledge with thanks DSR for technical and financial support.
182 This paper was also funded by KAKENHI 19H05270 and 17K00417.

References

[1] Gribkoff VK, Kaczmarek LK. The need for new approaches in CNS drug discovery: Why drugs have failed, and what can be done to improve outcomes. Neuropharmacology. 2017;120:11 – 19. Beyond Small Molecules for Neurological Disorders.

[2] Ransohoff RM. All (animal) models (of neurodegeneration) are wrong. Are they also useful? Journal of Experimental Medicine. 2018;215(12):2955–2958.

[3] Gordon J, Amini S, White MK. In: Amini S, White MK, editors. General Overview of Neuronal Cell Culture. Totowa, NJ: Humana Press; 2013. p. 1–8.

[4] Habib R, Noureen N, Nadeem N. Decoding Common Features of Neurodegenerative Disorders: From Differentially Expressed Genes to Pathways. Current Genomics. 2018;19(4):300–312.

[5] Taguchi YH. Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach. Unsupervised and Semi-Supervised Learning. Springer International Publishing; 2019. In press.

[6] Taguchi Yh. Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data. BMC Bioinformatics. 2019 Feb;19(13):388.

[7] Taguchi YH, Ng KL. Tensor Decomposition–Based Unsupervised Feature Extraction for Integrated Analysis of TCGA Data on MicrRNA Expression and Promoter Methylation of Genes in Ovarian Cancer. In: 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE); 2018. p. 195–200.

[8] Taguchi YH. Tensor Decomposition-Based Unsupervised Feature Extraction Can Identify the Universal Nature of Sequence-Nonspecific Off-Target Regulation of mRNA Mediated by MicroRNA Transfection. Cells. 2018;7(6):54.

[9] Taguchi YH. Tensor decomposition/principal component analysis based unsupervised feature extraction applied to brain gene expression and methylation profiles of social insects with multiple castes. BMC Bioinformatics. 2018;19(Suppl 4):99.

[10] Taguchi YH. One-class Differential Expression Analysis using Tensor Decomposition-based Unsupervised Feature Extraction Applied to Integrated Analysis of Multiple Omics Data from 26 Lung Adenocarcinoma

Cell Lines. In: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE); 2017. p. 131–138.

[11] Taguchi YH. Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing. PLoS ONE. 2017;12(8):e0183933.

[12] Taguchi YH. Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets. Sci Rep. 2017;7(1):13733.

[13] Taguchi YH. Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases. BMC Med Genomics. 2017;10(Suppl 4):67.

[14] Taguchi YH. Identification of Candidate Drugs for Heart Failure Using Tensor Decomposition-Based Unsupervised Feature Extraction Applied to Integrated Analysis of Gene Expression Between Heart Failure and DrugMatrix Datasets. In: Intelligent Computing Theories and Application. Springer International Publishing; 2017. p. 517–528.

[15] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

[16] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016 05;44(W1):W90–W97.

[17] Frigerio CS, Wolfs L, Fattorelli N, Thrupp N, Voytyuk I, Schmidt I, et al. The Major Risk Factors for Alzheimer's Disease: Age, Sex, and Genes Modulate the Microglia Response to Aβ Plaques. Cell Reports. 2019;27(4):1293 – 1306.e6.

[18] Leggio GM, Catania MV, Puzzo D, Spatuzza M, Pellitteri R, Gulisano W, et al. The antineoplastic drug flavopiridol reverses memory impairment induced by Amyloidβ_{1-42} oligomers in mice. Pharmacological Research. 2016;106:10 – 20.

[19] Hein LK, Apaja PM, Hattersley K, Grose RH, Xie J, Proud CG, et al. A novel fluorescent probe reveals starvation controls the commitment of amyloid precursor protein to the lysosome. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2017;1864(10):1554 – 1565.

[20] Heuvel CVD, Donkin JJ, Finnie JW, Blumbergs PC, Kuchel T, Koszyca B, et al. Downregulation of Amyloid Precursor Protein (APP) Expression following Post-Traumatic Cyclosporin-A Administration. Journal of Neurotrauma. 2004;21(11):1562–1572. PMID: 15684649.

[21] Eisele YS, Baumann M, Klebl B, Nordhammer C, Jucker M, Kilger E. Gleevec Increases Levels of the Amyloid Precursor Protein Intracellular Domain and of the Amyloid-egrading Enzyme Neprilysin. Molecular Biology of the Cell. 2007;18(9):3591–3600. PMID: 17626163.

[22] Litzenburger UM, Buenrostro JD, Wu B, Shen Y, Sheffield NC, Kathiria A, et al. Single-cell epigenomic variability reveals functional cancer heterogeneity. Genome Biology. 2017 Jan;18(1):15.

[23] Yuan D, Tao Y, Chen G, Shi T. Systematic expression analysis of ligand-receptor pairs reveals important cell-to-cell interactions inside glioma. Cell Communication and Signaling. 2019 May;17(1):48.

Table 1: Top ranked 10 compounds listed in "LINCS L1000 Chem Pert up" category in Enrichr. Overlap is that between selected 401 genes and genes selected in individual experiments.

Term	Overlap	P-value	Adjusted P-value
LJP006_HCC515_24H-alvocidib-10	28/221	7.99×10^{-15}	2.21×10^{-10}
LJP006_HCC515_24H-AZD-8055-10	24/188	5.87×10^{-13}	8.13×10^{-9}
LJP009_PC3_24H-CGP-60474-3.33	25/217	1.99×10^{-12}	1.14×10^{-8}
LJP005_MDAMB231_24H-AS-601245-10	20/132	2.05×10^{-12}	1.14×10^{-8}
LJP009_PC3_24H-saracatinib-10	24/196	1.47×10^{-12}	1.14×10^{-8}
LJP006_HCC515_24H-CGP-60474-0.37	24/225	2.89×10^{-11}	1.14×10^{-7}
LJP009_PC3_24H-PF-3758309-10	23/212	5.33×10^{-11}	1.84×10^{-7}
LJP005_HCC515_24H-WZ-3105-3.33	20/144	1.07×10^{-11}	4.95×10^{-8}
LJP006_HEPG2_24H-AZD-5438-10	21/182	1.17×10^{-10}	3.24×10^{-7}
LJP006_HCC515_24H-A443654-10	22/203	1.44×10^{-10}	3.62×10^{-7}

Table 2: Top ranked 10 compounds listed in "DrugMatrix" category in Enrichr. Overlap is that between selected 401 genes and genes selected in individual experiments.

Term	Overlap	P-value	Adjusted P-value
Cyclosporin_A-350_mg/kg_in_Corn_Oil-Rat-Bone_marrow-5d-up	51/315	2.26×10^{-31}	1.78×10^{-27}
Isoprenaline-4.2_mg/kg_in_Saline-Rat-Heart-5d-up	49/304	4.55×10^{-30}	1.79×10^{-26}
Hydroxyurea-400_mg/kg_in_Saline-Rat-Bone_marrow-5d-up	46/307	7.54×10^{-27}	1.49×10^{-23}
Netilmicin-40_mg/kg_in_Saline-Rat-Kidney-28d-up	45/314	1.90×10^{-25}	1.50×10^{-22}
Cyclosporin_A-350_mg/kg_in_Corn_Oil-Rat-Bone_marrow-3d-up	45/312	1.45×10^{-25}	1.42×10^{-22}
Chlorambucil-0.6_mg/kg_in_Corn_Oil-Rat-Spleen-0.25d-up	47/314	2.13×10^{-27}	5.60×10^{-24}
Tobramycin-40_mg/kg_in_Saline-Rat-Kidney-28d-up	45/311	1.26×10^{-25}	1.42×10^{-22}
Gemcitabine-11_mg/kg_in_Saline-Rat-Bone_marrow-3d-up	47/344	1.27×10^{-25}	1.42×10^{-22}
Terbutaline-130_mg/kg_in_Corn_Oil-Rat-Heart-3d-up	45/321	4.89×10^{-25}	2.41×10^{-22}
Cyclosporin_A-70_mg/kg_in_Corn_Oil-Rat-Bone_marrow-3d-up	45/320	4.28×10^{-25}	2.25×10^{-22}

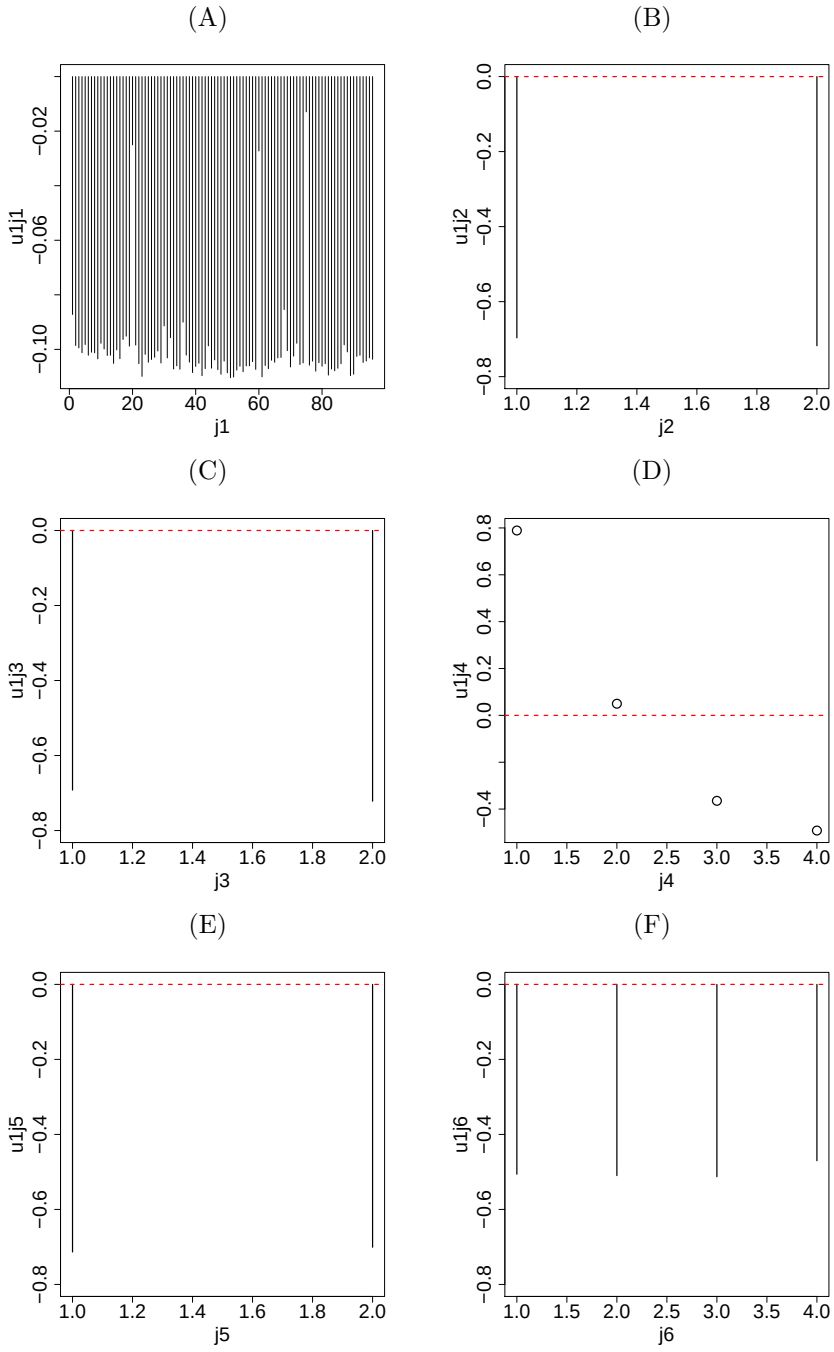

Figure 1: Singular value vectors. (A) u_{1j_1} (B) u_{1j_2}(C) u_{1j_3}(D) u_{2j_4}(E) u_{1j_5} (F) u_{1j_6}.

Table 3: Top ranked 10 compounds listed in "Drug Perturbations from GEO up" category in Enrichr. Overlap is that between selected 401 genes and genes selected in individual experiments.

Term	Overlap	P-value	Adjusted P-value
imatinib DB00619 mouse GSE51698 sample 2522	81/288	$2.27\times^{-70}$	$2.05\times^{-67}$
bleomycin DB00290 mouse GSE2640 sample 2851	80/329	$6.09\times^{-64}$	$2.75\times^{-61}$
soman 7305 rat GSE13428 sample 2640	86/532	$3.87\times^{-53}$	$3.50\times^{-51}$
coenzyme Q10 5281915 mouse GSE15129 sample 3464	76/302	$6.84\times^{-62}$	$2.06\times^{-59}$
N-METHYLFORMAMIDE 31254 rat GSE5509 sample 3570	70/283	$2.39\times^{-56}$	$3.60\times^{-54}$
Calcitonin 16132288 mouse GSE60761 sample 3446	65/220	$8.51\times^{-58}$	$1.92\times^{-55}$
cyclophosphamide 2907 mouse GSE2254 sample 3626	78/413	$2.47\times^{-53}$	$2.48\times^{-51}$
Calcitonin 16132288 mouse GSE60761 sample 3447	59/177	$5.88\times^{-56}$	$7.59\times^{-54}$
PRISTANE 15979 mouse GSE17297 sample 3229	71/291	$1.03\times^{-56}$	$1.87\times^{-54}$
coenzyme Q10 5281915 mouse GSE15129 sample 3456	76/396	$1.79\times^{-52}$	$1.35\times^{-50}$