Application of tensor decomposition based unsupervised feature extraction to bioinformatics/テンソル分解を用いた教師なし学習による変数選択のバイオインフォマティクスへの応用

Y-h. Taguchi/田口善弘

Department of Physics, Chuo University/中央大学物理学科,https://researchmap.jp/Yh_Taguchi/

Abstract

We have developed tensor decomposition based unsupervised feature extraction and applied it to various bioinformatics analysis. In the poster, we summarize from some mathematical basics to real applications.

Fundamental mathematics

A tensor, $x_{ijk} \in \mathbb{R}^{N\times M\times K}$, is the ith gene expression/promoter methylation/any other omics feature of the sample under the jth and the kth treatments . Using tensor decomposition (TD), x_{ijk} can be decomposed as

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

where $G(\ell_1,\ell_2,\ell_3) \in \mathbb{R}^{N\times M\times K}$ is a core tensor, $x_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $x_{\ell_2 j} \in \mathbb{R}^{M\times M}$, and $x_{\ell_3 k} \in \mathbb{R}^{K\times K}$ are singular value matrices that are supposed to be orthogonal matrices. Since x_{ijk} is as large as $G(\ell_1,\ell_2,\ell_3)$, it is obviously over complete. Using higher order singular value decomposition (HOSVD) algorithm, we can expect that summation in right hand side employing small number of Gs can well approximate x_{ijk}.
Feature selection was proposed by using the TD. Suppose that singular value matrices with two sets of $\{\ell_2'\}$ and $\{\ell_3'\}$ are expected to represent expected dependence upon treatments, e.g., tissue specificity, over expression under some treatments or enhanced drug responses. Then select a set of $\{\ell_1'\}$ associated with $G(\ell_1',\ell_2',\ell_3')$s with larger absolute values. This enables us to identify $x_{\ell_1' i}$s associated with expected treatments dependence specified above.

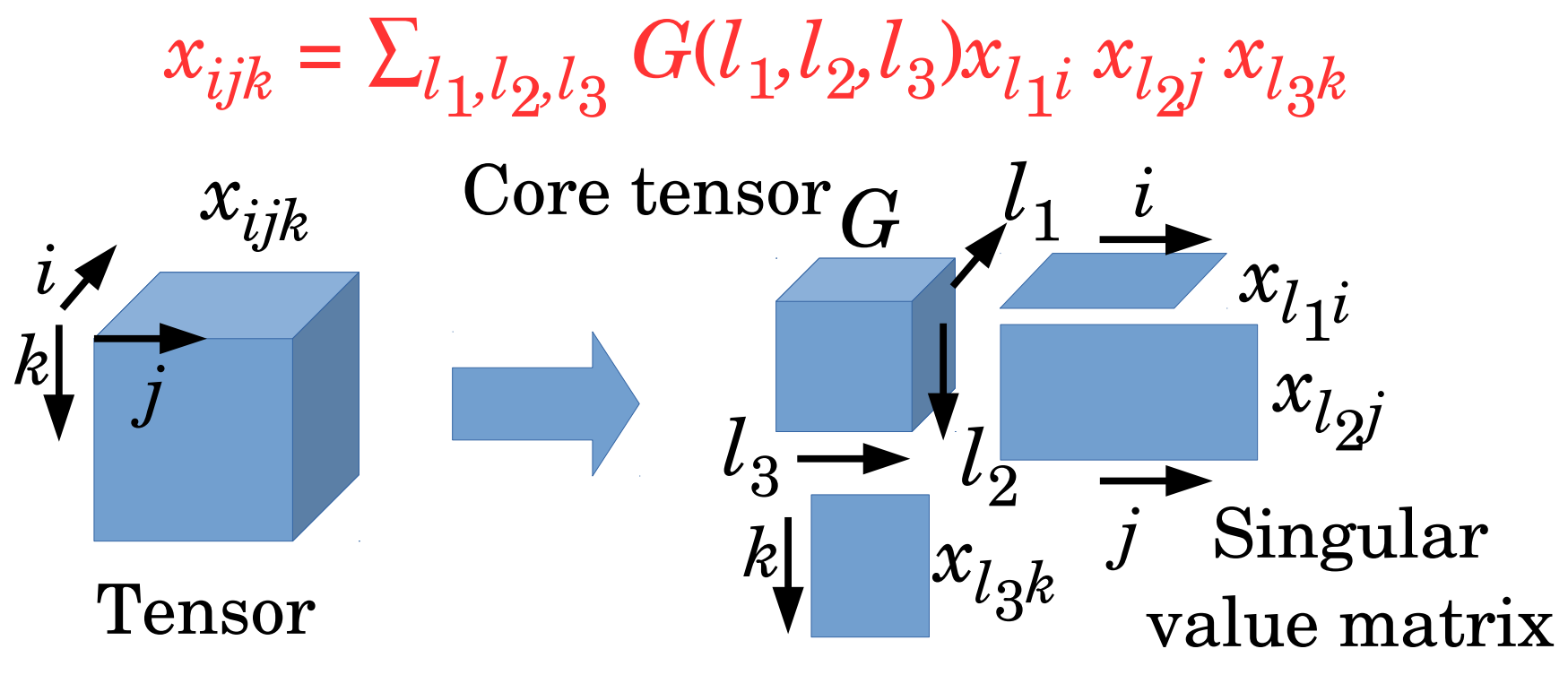

Figure 1: Schematic illustrates TD. $x_{ijk} \in \mathbb{R}^{N\times M\times K}$ is decomposed to product sum of $G(\ell_1,\ell_2,\ell_3) \in \mathbb{R}^{N\times M\times K}$, $x_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $x_{\ell_1 i} \in \mathbb{R}^{M\times M}$, and $x_{\ell_1 i} \in \mathbb{R}^{K\times K}$

In order to extract these genes, we assumed multiple Gaussian distribution for the selected $x_{\ell_1' i}$s. Then, P-values are attributed to the ith genes using χ^2 distribution.

$$P_i = P_{\chi^2}\left[> \sum_{\ell_1'} \left(\frac{x_{\ell_1' i}}{\sigma_{\ell_1'}} \right)^2 \right]$$

where $P_{\chi^2}[> x]$ is the cumulative probability assuming χ^2 distribution that the argument is larger than x and $\sigma_{\ell_1'}$ is standard deviation. If corrected P-values using multiple comparison comparison adjustment are less than 0.01, these genes are identified to share ordered common gene expression. TD was repeated using only selected genes when we need feature extraction (FE).

Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases

Although post-traumatic stress disorder (PTSD) is primarily a mental disease, it can also induce other disorders in remote tissues. The examples include cardiovascular, respiratory, musculoskeletal, neurological, and gastrointestinal disorders, diabetes, chronic pain, sleep disorders and other immune-mediated disorders.

Table 1: Samples used in this study. Numbers before/after comma are control/treated samples. h: hours, w: weeks.

stress	5 days		10 days			5 days		10 days	
RP	24h	1.5 w	24h	6w		24h	1.5 w	24h	6w
AY	3,2	5,4	3,4	3,4	HC	3,5	4,5	5,4	4,5
MPFC	4,5	5,5	3,4	4,4	SE	3,2	2,3	3,3	3,3
ST	5,5	5,5	5,4	4,4	VS	5,5	5,5	3,4	5,4
BLD	5,5	5,5	4,5	4,5	HT	5,5	4,5	5,5	5,5
HB	5,5	4,5	5,5	5,5	SP	5,5	5,5	5,4	5,5

RP: rest period, AY: amygdala, HC: hippocampus, MPFC: medial prefrontal cortex, SE: septal nucleus, ST: striatum, VS: ventral striatum, BLD: blood, HT: heart, HB: hemibrain, SP: spleen.

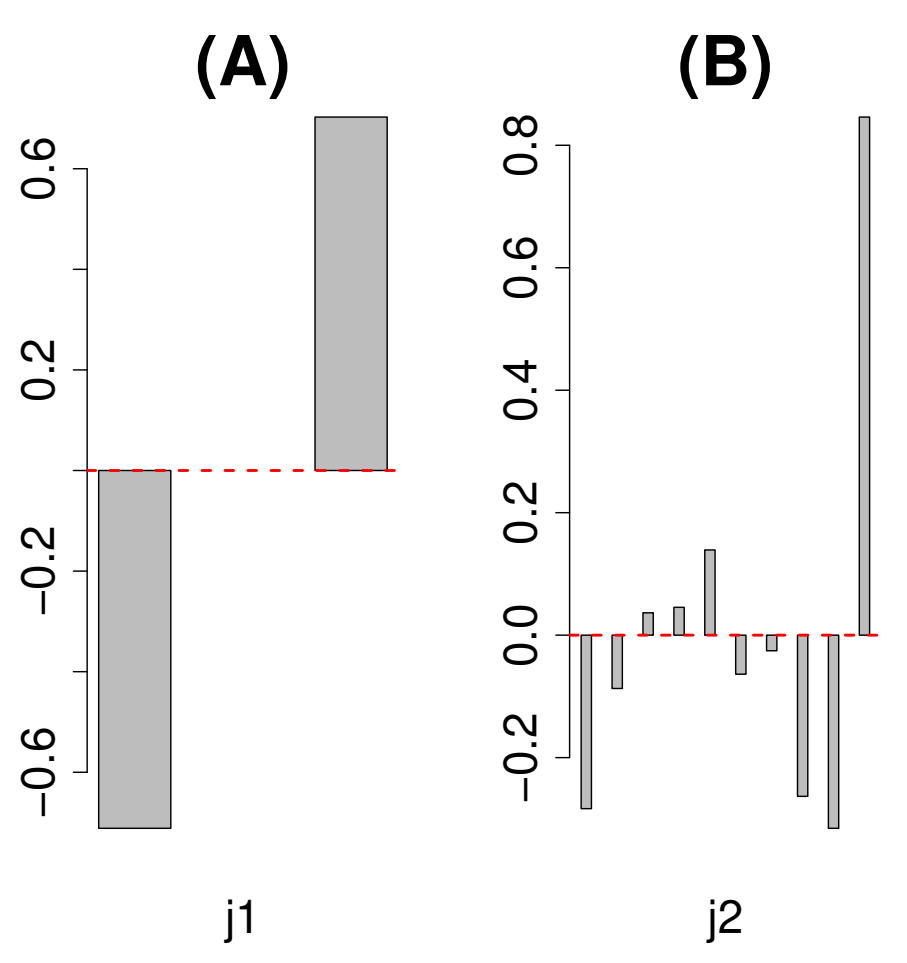

Figure 2: Singular value vectors employed (A) The second control-related or treatment-related singular value vector, $x_{\ell_1=2,j_1}$. Control: $j_1 = 1$, and treatment (stress): $j_1 = 2$. (B) The fourth tissue singular value vector, $x_{\ell_2=4,j_2}$, AY: $j_2 = 1$, HC: $j_2 = 2$, heart: $j_2 = 8$, hemibrain: $j_2 = 9$, and spleen: $j_2 = 10$.

We [1] applied TD based unsupervised FE to gene expression of various stressed mouse tissues with changing stress conditions (Table 1). Gene expression profiles were formatted as a tensor, $x_{i,j_1,j_2,j_3,j_4} \in \mathbb{R}^{N\times 2\times 10\times 2\times 3}$, of the ith probe, subjected to j_1th treatment ($j_1 = 1$: control, $j_1 = 2$: treated [stress-exposed] samples), in the j_2th tissue [$j_2 = 1$: amygdala (AY), $j_2 = 2$: hippocampus (HC), $j_2 = 3$: medial prefrontal cortex (MPFC), $j_2 = 4$: septal nucleus (SE), $j_2 = 5$: striatum (ST), $j_2 = 6$: ventral striatum (VS), $j_2 = 7$: blood, $j_2 = 8$: heart, $j_2 = 9$: hemibrain, $j_2 = 10$: spleen], with the j_3th stress duration ($j_3 = 1$: 10 days, $j_3 = 2$: five days) and j_4th rest period after application of stress ($j_4 = 1$: 1.5 weeks, $j_4 = 2$: 24 hours, $j_4 = 3$: 6 weeks). Zero values were assigned to missing observations (e.g., measurements at 6 weeks after a 5-day period of stress are not available).
After applying TD to x_{i,j_1,j_2,j_3,j_4} as

$$x_{i,j_1,\ldots,j_4} = \sum_{\ell_1,\ldots,\ell_5} G(\ell_1,\ldots,\ell_5)\cdot x_{\ell_5,i} \prod_{k=1}^{4} x_{\ell_k,j_k}$$

we found that $\ell_1 = 2$ represents distinction between treated and control conditions and $\ell_2 = 4$ represents co-expression among brain subregions and heart. Then we tried to find which $G(\ell_1 = 2, \ell_2 = 4, \ell_3, \ell_4, \ell_5)$s have larger absolute values (Table 2). It is obvious that $\ell_5 = 1, 4, 11$ are associated with Gs with larger absolute values.

Table 2: Top-ranked $G(\ell_1 = 2, \ell_2 = 4, \ell_3, \ell_4, \ell_5)$ with greater absolute values.

ℓ_3	ℓ_4	ℓ_5	$G(2,4,\ell_3,\ell_4,\ell_5)$
1	1	11	-35.0
1	1	1	-30.8
2	2	1	-30.3
2	3	4	-30.0
2	3	1	28.7
2	2	4	28.5

After that, 801 probes associated with adjusted P-values less than 0.01 were selected as outliers using these three gene singular value vectors. Genes associated with these identified 801 probes were validated and turned out to be biologically reliable (see Ref. [1] for more details).

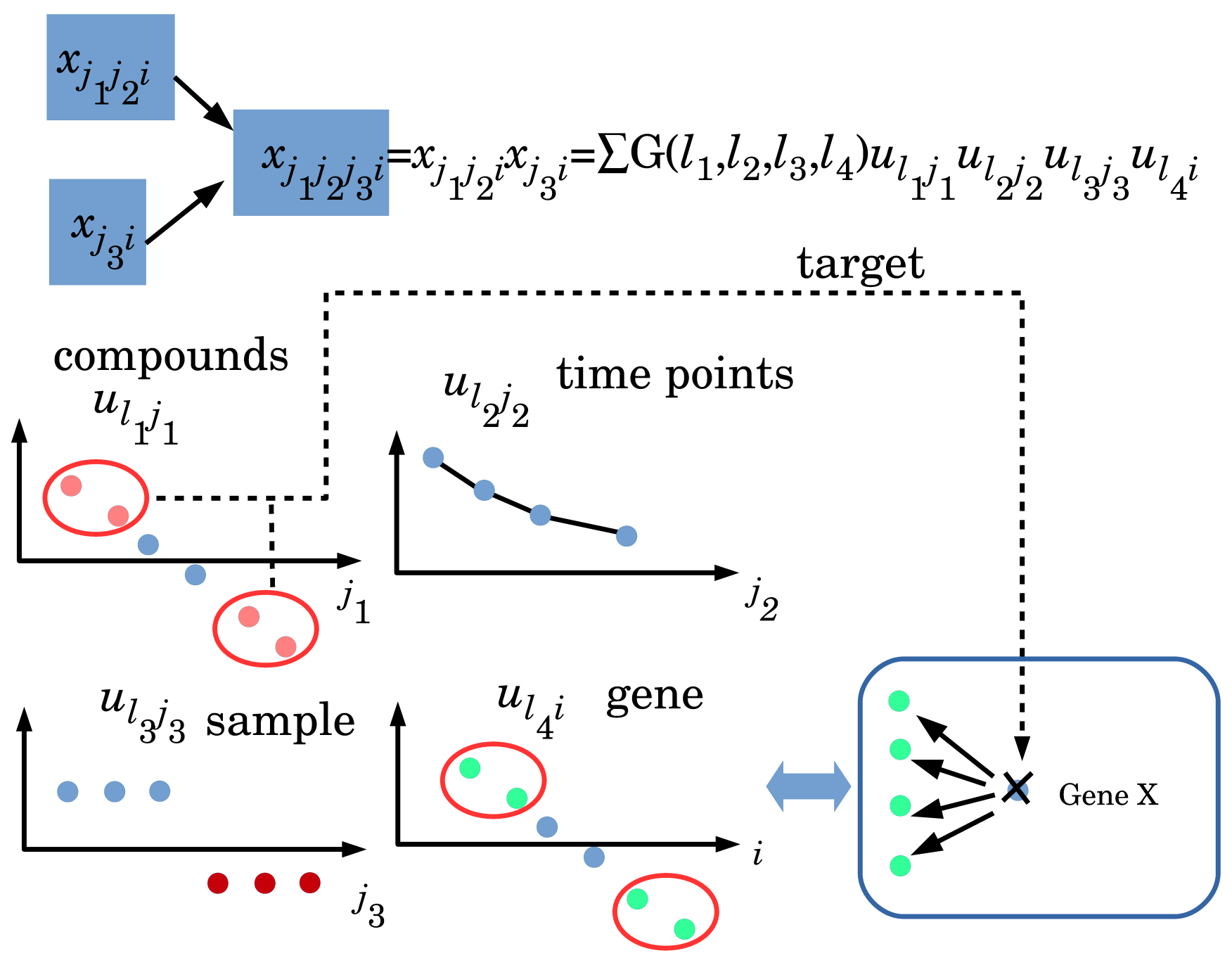

Figure 3: Intuitive illustration of the present strategy.

Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets

In the next application of TD based unsupervised FE, we aimed *in silico* drug discovery[2].
Suppose there is a tensor, $x_{j_1 j_2 i}$, which represents the ith gene expression at the j_2th time point after the j_1th compound is given to a rat; these data are taken from the DrugMatrix dataset (Fig. 3). There is also a matrix, $x_{j_3 i}$, which represents the ith gene expression of the j_3th sample; samples typically include disease samples and control samples. Tensor $\tilde{x}_{j_1 j_2 j_3 i}$ was generated as a 'mathematical product' of $x_{j_1 j_2 i}$ and $x_{j_3 i}$. Then, tensor $\tilde{x}_{j_1 j_2 j_3 i}$ is decomposed, and singular value matrix of compounds $u_{\ell_1 j_1}$, singular value matrix of time points $u_{\ell_2 j_2}$, sample singular value matrix $u_{\ell_3 j_3}$, and gene singular value matrix $u_{\ell_4 i}$ are obtained. Among them, I selected the combinations of $\ell_k, 1 \le k \le 4$, which are simultaneously associated with all of the following: i) core tensor $G(\ell_1,\ell_2,\ell_3,\ell_4)$ with a large enough absolute value, ii) a singular value vector of time points, $u_{\ell_2 j_2}$, whose value significantly varies with time, and iii) sample singular value vector $u_{\ell_3 j_3}$. These parameters are different between a disease (red filled circles) and control samples (cyan filled circles). Finally, using gene singular value vector $u_{\ell_4 i}$ and compound singular value vector $u_{\ell_1 j_1}$, compounds (filled pink circles) and genes (filled light-green circles) associated with $G(\ell_1,\ell_2,\ell_3,\ell_4)$s with large enough absolute values are selected. Next, if the selected genes are coincident with the genes associated with a significant alteration when gene X is knocked out (or overexpressed), then the compounds are assumed to target gene X.

Table 3: Fisher's exact test (P_F) and the uncorrected χ^2 test (P_{χ^2}) of known drug target proteins regarding the inference of the present study. Rows: known drug target proteins (DINIES). Columns: Inferred drug target proteins using 'Single Gene Perturbations from GEO up' or 'Single Gene Perturbations from GEO down'. OR: odds ratio

		Single Gene Perturbations from GEO up					Single Gene Perturbations from GEO down				
		F	T	P_F	P_{χ^2}	RO	F	T	P_F	P_{χ^2}	RO
heart failure	F	521	517	3.4×10^{-4}	3.9×10^{-4}	3.02	628	416	1.3×10^{-3}	7.3×10^{-4}	2.61
	T	13	39				19	33			
PTSD	F	500	560	3.8×10^{-2}	3.1×10^{-2}	2.67	532	529	6.1×10^{-3}	4.5×10^{-3}	3.81
	T	6	18				5	19			
ALL	F	979	89	2.7×10^{-1}	3.0×10^{-1}	2.19	1009	57	1.0×10^{0}	-	-
	T	10	2				12	0			
diabetes	F	889	177	1.2×10^{-2}	7.1×10^{-3}	3.00	936	130	3.6×10^{-4}	2.0×10^{-5}	5.13
	T	15	9				14	10			
renal carcinoma	F	847	219	2.0×10^{-2}	1.2×10^{-2}	2.75	895	169	4.3×10^{-2}	2.2×10^{-2}	2.64
	T	14	10				16	8			
cirrhosis	F	572	219	1.1×10^{-2}	8.1×10^{-3}	2.91	595	169	1.6×10^{-3}	1.1×10^{-3}	3.81
	T	8	10				7	8			

Table 3 shows the result. For five diseases other than ALL, identified combinations of drugs and target proteins are significantly overlapped with known pairs of drugs and target proteins.

Other applications

We have applied TD based unsupervised FE to other various applications. More advanced discussions about the methods used for drug discovery [2, 3] is found in Ref. [4]. Applications of TD based unsupervised FE to integrated analysis of multi-omics data set is found in Ref. [5]. Genetic and epigenetic background and miRNA transfection mediated sequence non-specific side effect are also discussed in APBC2018 [6] and GIW2017 [7], respectively. Correlation between miRNA expressin and DNA methylation was studied [8], too.

References

[1] Y-.H. Taguchi.
Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases.
BMC Med. Genomics, 10(S4):67, 2017.

[2] 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):13733, oct 2017.

[3] 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, 2019.
in press.

[4] Y h. Taguchi.
Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing.
PLOS ONE, 12(8):e0183933, aug 2017.

[5] Y h. Taguchi.
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.
pages 131–138, 2017.
IEEE 17th International Conference on Bioinformatics and Bioengineering.

[6] Y.-H. Taguchi.
Tensor decomposition-based and principal-component-analysis-based unsupervised feature extraction applied to the gene expression and methylation profiles in the brains of social insects with multiple castes.
BMC Bioinformatics, 19(S4):99, may 2018.

[7] Y.-H. Taguchi.
Tensor decomposition-based unsupervised feature extraction can identify the universal nature of sequence-nonspecific off-target regulation of mRNA mediated by MicroRNA transfection.
Cells, 7(6):54, jun 2018.

[8] Y h. Taguchi and Ka-Lok Ng.
Tensor decomposition–based unsupervised feature extraction for integrated analysis of TCGA data on microRNA expression and promoter methylation of genes in ovarian cancer.
pages 195–300, 2018.
IEEE 18th International Conference on Bioinformatics and Bioengineering, in press.

Acknowledgements

These studies are supported by KAKENHI 17K00417.