Chapter 7
Application of TD Based Unsupervised FE to Bioinformatics

May my wish never come true.
Rikka Takarada, SSSS.GRIDMAN, Season 1, Episode 12

7.1 Introduction

Because of continuous price reduction of multiomics data measurements, including gene expression, promoter methylation, SNP, histone modification, and miRNA expression, more number of experimental conditions come to be considered. For example, if gene expression is measured for various tissues of patients, gene expression has better to be formatted, not in matrix, but in tensor, as patients vs tissue vs genes. In this case, TD rather than PCA is a suitable technology to apply. On the other hand, in the previous chapter, we aimed various integrated analysis, e.g., miRNA and mRNA expression, mRNA expression and methylation, mRNA expression of two species. If genes or features are shared in the integrated analysis, generation of case I or II tensor and application of TD to it is a suitable treatment. In the following, we introduce some application of TD based unsupervised FE to either of these cases.

7.2 PTSD Mediated Heart Diseases

The first example to be processed as tensor form is PTSD mediated heart diseases. Although this disease has already been analyzed in the previous chapter (Sect. 6.4.1), the data set analyzed there includes only one tissue, heart. Nonetheless, if one would like to understand how PTSD mediates heart disease, we need to know gene expression of both heart and brain. Fortunately, there is a such kind of data set. In this section, I would like to demonstrate the usefulness of TD based unsupervised FE applied to gene expression of multiple tissues aiming to understand PTSD mediated heart disease based upon the recent publication [24].

Y-h. Taguchi, *Unsupervised Feature Extraction Applied to Bioinformatics*,
Unsupervised and Semi-Supervised Learning,
https://doi.org/10.1007/978-3-030-22456-1_7

Table 7.1 Samples used in this study

Stress, days	5		10			5		10	
Rest period	24 h	1.5 w	24 h	6w		24 h	1.5 w	24 h	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
Blood	5,5	5,5	4,5	4,5	Heart	5,5	4,5	5,5	5,5
Hemibrain	5,5	4,5	5,5	5,5	Spleen	5,5	5,5	5,4	5,5

Numbers before/after comma are control/treated samples. *h* hours, *w* weeks, *AY* amygdala, *HC* hippocampus, *MPFC* medial prefrontal cortex, *SE* septal nucleus, *ST* striatum, *VS* ventral striatum

The data set analyzed is composed of the following samples (Table 7.1). It includes ten tissues under eight experimental conditions. This data set is formatted as a five-mode tensor, $x_{ij_1j_2j_3j_4} \in \mathbb{R}^{43699\times 2\times 10\times 2\times 3}$, 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, $j_3 = 2$: 5 days) and j_4th rest period after application of stress ($j_4 = 1$: 1.5 weeks, $j_4 = 2$: 24 h, $j_4 = 3$: 6 weeks). Zero values are assigned to missing observations (e.g., measurements at 6 weeks after a 5-day period of stress are not available).

HOSVD algorithm (Fig. 3.8) is applied to $x_{ij_1j_2j_3j_4}$ as

$$x_{ij_1j_2j_3j_4} = \sum_{\ell_5=1}^{43699}\sum_{\ell_1=1}^{2}\sum_{\ell_2=1}^{10}\sum_{\ell_3=1}^{2}\sum_{\ell_4=1}^{3} G(\ell_1,\ell_2,\ell_3,\ell_4,\ell_5)u^{(j_1)}_{\ell_1 j_1}u^{(j_2)}_{\ell_2 j_2}u^{(j_3)}_{\ell_3 j_3}u^{(j_4)}_{\ell_4 j_4}u^{(i)}_{\ell_5 i} \tag{7.1}$$

where $u^{(i)}_{\ell_5 i} \in \mathbb{R}^{43699\times 43699}$, $u^{(j_1)}_{\ell_1 j_1} \in \mathbb{R}^{2\times 2}$, $u^{(j_2)}_{\ell_2 j_2} \in \mathbb{R}^{10\times 10}$, $u^{(j_3)}_{\ell_3 j_3} \in \mathbb{R}^{2\times 2}$, and $u^{(j_4)}_{\ell_4 j_4} \in \mathbb{R}^{3\times 3}$ are singular value vectors and $G(\ell_1,\ell_2,\ell_3,\ell_4,\ell_5) \in \mathbb{R}^{43699\times 2\times 10\times 2\times 3}$ is a core tensor.

We need to specify which singular value vector attributed to genes, $\boldsymbol{u}^{(i)}_{\ell_1}$, is used for gene selection. For this purpose, we investigate other singular value vectors, $\boldsymbol{u}^{(j_k)}_{\ell_k}$, $1 \le k \le 4$. One of the important points is tissue specificity. What I would like to find is a set of genes expressive in common between heart and brain. Because $1 \le j \le 6$ and $j = 9$ correspond to brain and $j = 8$ corresponds to heart, we need to find $\boldsymbol{u}^{(j_2)}_{\ell_2}$ expressive in common $j = 1, 2, \cdots, 6, 8, 9$. Figure 7.1 shows the singular value vectors, $\boldsymbol{u}^{(j_2)}_{\ell_2}$, $1 \le \ell_2 \le 10$. Although no $\boldsymbol{u}^{(j_2)}_{\ell_2}$ fully satisfies this requirement, $\boldsymbol{u}^{(j_2)}_{4}$ relatively fulfills this requirement. $\boldsymbol{u}^{(j_2)}_{4}$ are negatively signed in common for $j = 1, 2, 8, 9$ that correspond to AY, HC, heart, and hemibrain. Especially, because AY and HC are very important in PTSD [14], it is promising that we can get singular value vector expressive in common AY, HC, and heart.

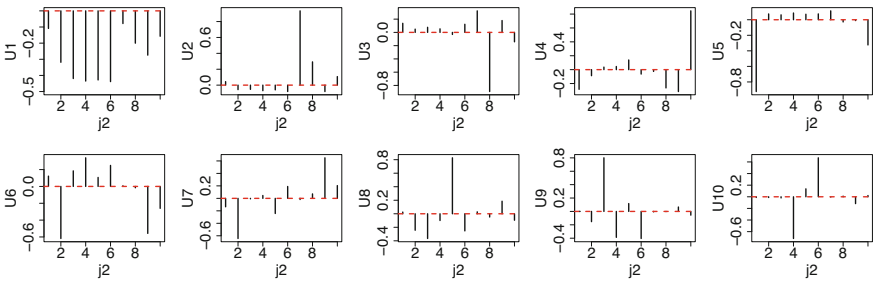

Fig. 7.1 Singular value vectors, $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \leq \ell_2 \leq 10$. Red horizontal broken lines show baseline

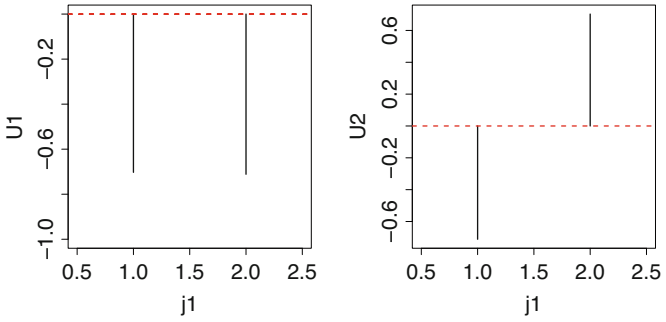

Fig. 7.2 Singular value vectors, $\boldsymbol{u}_{\ell_1}^{(j_1)}$, $\ell_1 = 1, 2$. Red horizontal broken lines show baseline

The next important requirement is that control and stressed samples should be oppositely expressive. This means, $u_{\ell_1 1}^{(j_1)} = -u_{\ell_1 2}^{(j_1)}$. This requirement is easy to fulfill because $u_{\ell_1 1}^{(j_1)} = -u_{\ell_1 2}^{(j_1)}$ or $u_{\ell_1 1}^{(j_1)} = u_{\ell_1 2}^{(j_1)}$ must be satisfied when there are only two classes and mean is zero. Figure 7.2 shows the singular value vectors, $\boldsymbol{u}_{\ell_1}^{(j_1)}$, $\ell_1 = 1, 2$. As expected, $\ell_1 = 2$ corresponds to the reversed sign between control and stressed samples.

Because there are no known pre-defined desirable properties for experimental conditions, i.e. stress and rest period, we should find $G(2, 4, \ell_3, \ell_4, \ell_5)$ with the larger absolute values. Table 7.2 shows the top ranked G with larger absolute values. Then we can find that $\ell_5 = 1, 4, 11$ are associated with $G(2, 4, \ell_3, \ell_4, \ell_5)$ with the larger absolute values. Thus we decided to attribute P-values using $\ell_5 = 1, 4, 11$ with assuming χ^2 distribution as

$$P_i = P_{\chi^2}\left[> \sum_{\ell_5 = 1,4,11} \left(\frac{u_{\ell_5 i}}{\sigma_{\ell_5}} \right)^2 \right]. \tag{7.2}$$

P-values are corrected by BH criterion and 801 probes associated with adjusted P-values less than 0.01 are selected.

Table 7.2 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

Table 7.3 Thirteen combinations of tissues and experimental conditions where the selected 801 probes are differentially expressed between stress-exposed and control samples

Stress duration	10 days		5 days	
Rest period	24 h	6 weeks	24 h	1.5 weeks
AY		◯		◯
HC		◯	◯	◯
MPFC		◯		
Heart	◯			◯
Hemibrain			◯	◯
Spleen		◯	◯	◯

MPFC: medial prefrontal cortex. ◯: associated with P-values that are computed by t test, adjusted by BH criterion and less than 0.01

The first validation of selected 801 probes is to see if these are expressed distinctly between control and stressed samples, selectively on only heart and brain. In order to confirm this, we apply t test to the selected 801 probes between control and stressed samples for all combination of tissues, rest and stressed period. P-values are corrected by BH criterion and conditions associated with adjusted P-values less than 0.01 are considered to be expressed distinctly and significantly between control and stressed samples. Table 7.3 shows the results. The selected 801 genes are expressed distinctly between control and stressed samples, selectively in heart, HC, and AY (it is also in spleen, because it is oppositely expressed toward heart, HC, and AY as shown in Fig. 7.1).

Here we would like to emphasize the difficulty of gene selection in this data set. As mentioned above, what we are aiming is quite abstract, i.e., "genes expressive in common between brain and heart as well as distinctly between control and stressed samples." As a result, we realize that common expression between AY, HC, and heart is possible (with the investigation of $\boldsymbol{u}_4^{(j_2)}$ in Fig. 7.1). Generally, it is impossible to know this combination in advance. When no clear purpose is given in advance, supervised methods cannot perform well while unsupervised methods can.

In order to see how well other conventional supervised methods perform, we test three methods, SAM, limma, and categorical regression analysis. The first example to be compared with TD based unsupervised FE is categorical regression analysis. For the data set shown in Table 7.1, the only possible way to apply categorical regression is to treat it as 80 classes (10 tissues vs four experimental conditions vs control and stressed samples). Although it is better to consider the pair of control and stressed samples, it is impossible. Typically, although ratio might be taken, because

Table 7.4 Results of gene selection based on categorical regression

Adjusted P-values	$P > 0.01$	$P < 0.01$	$P > 0.05$	$P < 0.05$	$P > 0.1$	$P < 0.1$
Number of probes	2222	41,157	1986	41,713	1839	41,860

P-values are adjusted by BH criterion

Table 7.5 Results by SAM

	Delta	p0	False	Called	FDR
1	0.1	0.011	38,538.08	43,379	0.0094
2	11.4	0.011	0.02	5424	3.9e−08
3	22.7	0.011	0	323	0
4	34.0	0.011	0	40	0
5	45.2	0.011	0	7	0
6	56.5	0.011	0	4	0
7	67.8	0.011	0	2	0
8	79.1	0.011	0	1	0
9	90.3	0.011	0	1	0
10	101.6	0.011	0	1	0

p0 is the ratio of the null hypothesis, FDR corresponds to the adjusted P-values. Called is the number of genes that break the null hypothesis. Expected number of false positives is False $\times$ FDR $\times$ p0

it is not paired samples, i.e., there is no one-to-one correspondence, we cannot take ratio. Table 7.4 shows the result of categorical regression analysis. Because of treatment as 80 classes, genes associated with any kind of distinction are detected (i.e., associated with significantly small adjusted P-values). As a result, almost all genes are judged as distinct between some combinations. It is obvious that this result is not desirable for our purpose, "genes expressive in common between brain and heart distinctly between control and stressed samples," at all, because of lack of specificity. To screen these genes, we need some additional criterion that TD based unsupervised FE does not require. Thus, TD based unsupervised is more fitted to the present purpose than categorical regression.

Next, we apply SAM with assuming 80 classes to the data set shown in Table 7.1. Table 7.5 shows the result of SAM. $p0$, which represents the contribution of null hypothesis that no distinction exist among 80 classes, is 1%. This means, almost all genes are distinctly expressive in either of these combinations. Although FDR corresponds to the adjusted P-values, it is clear that all genes are associated with FDR less than 0.01. Although this conclusion itself is coincident with that of categorical regression, in this sense SAM is not useful to select "genes expressive in common between brain and heart distinctly between control and stressed samples," either.

Finally, we apply limma to the data set shown in Table 7.1. Fortunately, limma enables us to select genes that are distinct between any pairs of controls and samples. Thus, we apply limma in two ways. One assumes 80 classes (case A in Table 7.6) and the other assumes 40 classes (case B in Table 7.6) composed of forty (10

Table 7.6 Results of gene selection based on limma

Adjusted P-values	$P > 0.01$	$P < 0.01$	$P > 0.05$	$P < 0.05$	$P > 0.1$	$P < 0.1$
	Case A : not considering differential expression					
Number of probes	0	43,379	0	43,379	0	43,379
	Case B: considering differential expression					
Number of probes	25,992	17,387	17,745	25,634	13,542	29,837

P-values are adjusted by limma itself

Table 7.7 KEGG pathway enrichment by the 457 genes identified by TD based unsupervised FE

Category	Term	Genes count	%	P-value	Adjusted P-value
KEGG_PATHWAY	Ribosome	57	12.8	8.4×10^{-58}	1.0×10^{-55}
KEGG_PATHWAY	Parkinson's disease	48	10.8	3.6×10^{-33}	2.2×10^{-31}
KEGG_PATHWAY	Oxidative phosphorylation	47	10.5	1.7×10^{-32}	6.9×10^{-31}
KEGG_PATHWAY	Alzheimer's disease	50	11.2	2.5×10^{-28}	7.5×10^{-27}
KEGG_PATHWAY	Huntington's disease	48	10.8	3.6×10^{-26}	8.6×10^{-25}
KEGG_PATHWAY	Cardiac muscle contraction	30	6.7	2.4×10^{-21}	4.8×10^{-20}
KEGG_PATHWAY	Glycolysis/ gluconeogenesis	10	2.2	1.5×10^{-3}	2.6×10^{-2}

Adjusted P-values are by BH criterion

tissues vs four experimental conditions) pairwise combinations between control and stress samples. Possibly because of its advanced feature, limma successfully denies the detection of genes expressive distinct among any pairs of 80 classes (case A). Nevertheless, limma still detects too many positives in 40 pairwise comparisons (case B). As expected, because of lack of well-defined screening criterion, three supervised methods are useless to find "genes expressive in common between brain and heart as well as distinctly between control and stressed samples." In conclusion, none of the three conventional supervised methods are as useful as TD based unsupervised FE for the present purpose.

Although TD based unsupervised FE successfully identifies genes expressive distinct between control and stressed samples in tissue specific manner (Table 7.3), if it is biologically useless, it cannot be considered to be successful. In order to evaluate selected probes biologically, we try to identify protein coding genes associated with these 801 probes. Then, we find 457 genes (because of lack of space, we cannot list all of 457 genes, which is available as Additional file 5 [24], if the readers are particularly interested in them). We upload 457 genes to DAVID. The result is quite promising. Table 7.7 shows the enriched KEGG pathway associated with adjusted P-values less than 0.05. They include four neurodegenerative diseases as well as one cardiac problem. Thus, they are quite suitable to be candidate genes that cause PTSD mediated heart diseases as those in Table 6.20 where PTSD mediated heart disease is investigated by PCA based unsupervised FE.

7.3 Drug Discovery From Gene Expression

Drug discovery is time-consuming and expensive processes. It starts from preparing as many small molecules as possible. Then, tries to find one effective to target diseases by exhausted search. The number of initially prepared molecule can be 10^4; testing this many number of compounds causes huge amount of money and long period. If we can reduce the number of initial candidate small molecules to one tenth, it benefits so much to reduce the time and cost required.

In this sense, the so-called in silico drug discovery develops with much expectation to fulfill this requirement. In silico drug discovery is aiming to identify candidate small molecules without *wet* experiments. With making full use of recently developed computational power, including CPU with high speed computing, huge storage that can store massive information as well as recently developed machine learning technique, in silico drug discovery enables us to prepare set of more promising candidate small molecules as drugs.

Traditionally, there are two main streams of in silico drug discovery. One is ligand-based drug design [1] (LBDD) and the other is structure-based drug design [3] (SBDD). LBDD is aiming to identify new candidate drug compounds based upon the similarity with known drugs. LBDD has huge varieties depending upon how similarity is defined. The advantage of LBDD is that it has more trust, i.e. larger probability to find true drug compounds, and requires smaller computational resources than SBDD. The disadvantage of LBDD is that it requires the information of known drugs and fails to find new drug candidates that lack similarity with known drug. On the contrary, SBDD has the advantage that it can predict new candidate drugs without the information of known drugs. The disadvantage of SBDD is that it requires massive computation, because it must execute docking simulation between drug candidate compounds and target proteins. Another disadvantage of SBDD is that it needs protein tertiary structure to which individual candidate drug compounds must bind. Experimental measurements of protein tertiary structure itself are difficult tasks. Although it has become much easier because of the invention of cryo-electron microscopy [10] than before, it still needs to pay much amount of money and time. When there are no protein tertiary structures available, protein tertiary structure itself must be computationally predicted [6]. The prediction inevitably has inaccuracy that affects the prediction of binding affinity of small molecules.

In order to compensate these disadvantages of LBDD and SBDD, the third option is recently proposed: drug design from gene expression [5]. Post-treatment gene expression can be used to screen candidate compounds for their ability to induce the target phenotype. This approach is very useful once post-treatment gene expression is available. In this section, we try to make use of TD based unsupervised FE to predict new drug target with analyzing post-treatment gene expression [27].

Post-treatment gene expression is obtained from LINCS [20]. L1000 is highly reproducible, comparable to RNA sequencing, and suitable for computational inference of the expression levels of 81% of non-measured transcripts. Gene expression

profile is available in GEO with GEO ID GSE70138. Table 7.8 summarizes the gene expression profiles. They include 13 cell lines to which 100–300 compounds (denoted as "all compounds") are treated. One problem of this data set is that it includes only 978 genes' expression profiles, because it is measured by Luminex scanners. Gene expression profiles in individual cell lines are formatted as tensor, $x_{ijk} \in \mathbb{R}^{978\times 6\times K}$; i denotes gene (probe), j denotes dose density of drug compound, and k stands for individual compounds among K total number of compounds that correspond to "all compounds" in Table 7.8. HOSVD algorithm (Fig. 3.8) is applied as

$$x_{ijk} = \sum_{\ell_1=1}^{978} \sum_{\ell_2=1}^{6} \sum_{\ell_3=1}^{K} G(\ell_1, \ell_2, \ell_3) u^{(i)}_{\ell_1 i} u^{(j)}_{\ell_2 i} u^{(k)}_{\ell_3 k} \tag{7.3}$$

where $\boldsymbol{u}^{(i)}_{\ell_1} \in \mathbb{R}^{978}$, $\boldsymbol{u}^{(j)}_{\ell_2} \in \mathbb{R}^{6}$, $\boldsymbol{u}^{(k)}_{\ell_3} \in \mathbb{R}^{K}$, are the singular value vectors, and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{978\times 6\times K}$ is a core tensor.

The first step is to identify genes whose expression is altered by drug treatment. In order that, we try to identify which $\boldsymbol{u}^{(j)}$ has monotonic dependence upon dose

Table 7.8 The number of the inferred compounds and inferred genes associated with significant dose-dependent activity

Cell lines	BT20	HS578T	MCF10A	MCF7	MDAMB231	SKBR3
Tumor	Breast					
Inferred genes	41	57	42	55	41	46
Inferred compounds	4	3	2	6	5	6
All compounds	110	106	106	108	108	106
Predicted targets	418	576	476	480	560	423

Cell lines	A549	HCC515	HA1E	HEPG2	HT29	PC3
Tumor	Lung		Kidney	Liver	Colon	Prostate
Inferred genes	45	46	48	54	50	63
Inferred compounds	8	5	7	2	2	9
All compounds	265	270	262	269	270	270
Predicted targets	428	352	423	396	358	439

Cell lines	A375
Tumor	Melanoma
Inferred genes	43
Inferred compounds	6
All compounds	269
Predicted targets	421

The target proteins predicted by means of the comparison with the data showing upregulation of the expression of individual genes ("predicted targets") are also shown

The full list of inferred genes and predicted targets is available in Additional file 7 [27]. Inferred compounds are presented in Table 7.9. "All compounds" rows represent the total number of compounds used for the treatment of each cell line

density. Figure 7.3 shows $\boldsymbol{u}_{\ell_2}^{(j)}, 1 \leq \ell_2 \leq 3$ for 13 cell lines listed in Table 7.8. It is obvious that $\boldsymbol{u}_2^{(j)}$ shows almost linear dependence upon dose independent of cell lines. The next task is to identify $G(\ell_1, 2, \ell_3)$ with larger absolute values in order

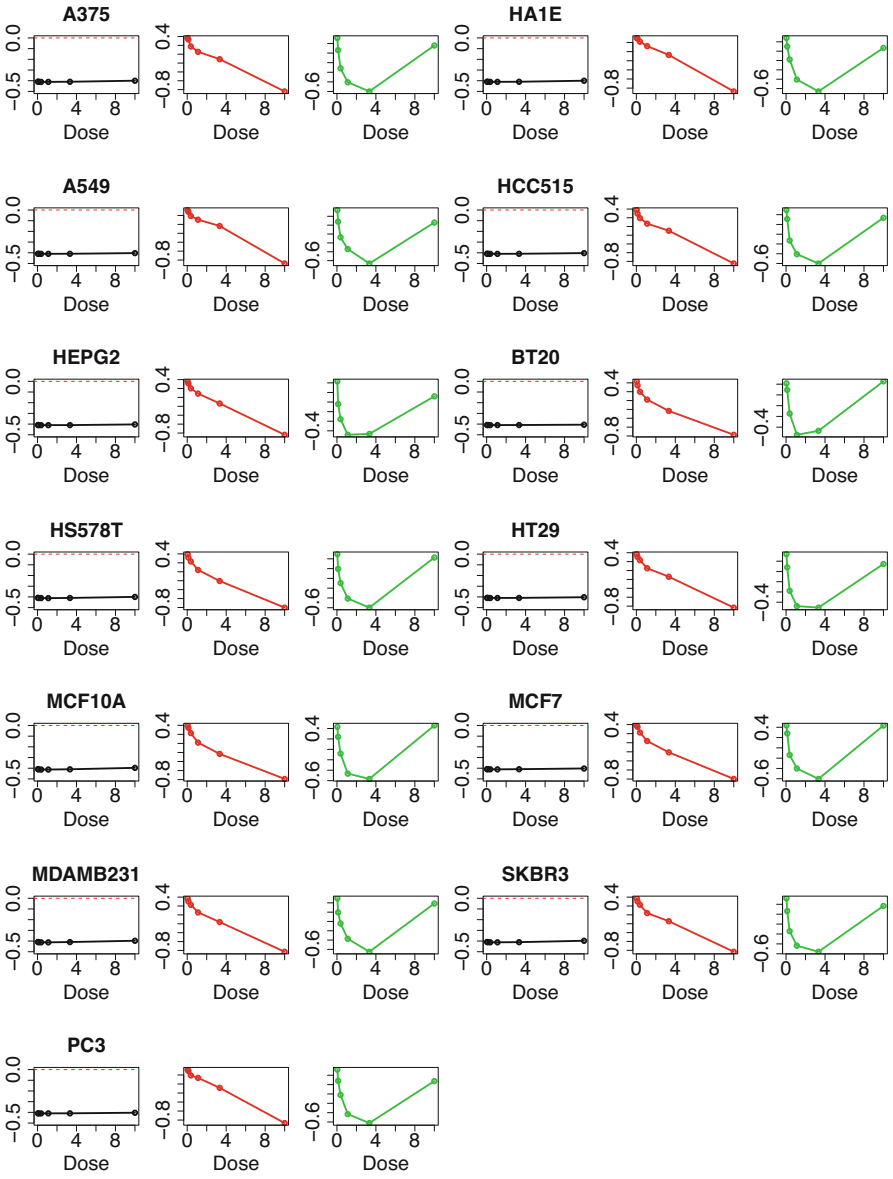

Fig. 7.3 Singular value vectors, $\boldsymbol{u}_{\ell_2}^{(j)}, 1 \leq \ell_2 \leq 3$. Red horizontal broken lines indicates baseline. Black: $\ell_2 = 1$, red: $\ell_2 = 2$, green: $\ell_2 = 3$

to decide which $\boldsymbol{u}_{\ell_1}^{(i)}$ and $\boldsymbol{u}_{\ell_3}^{(k)}$ are used for selecting the combinations of genes and compounds that commit linear dose dependence. Because

$$G(\ell_1 \leq 6, \ell_2 \leq 6, \ell_3 \leq 6) = \frac{\sum_{\ell_1 \leq 6, \ell_2 \leq 6, \ell_3 \leq 6} G(\ell_1, \ell_2, \ell_3)^2}{\sum_{\ell_1, \ell_2, \ell_3} G(\ell_1, \ell_2, \ell_3)^2} \tag{7.4}$$

exceeds 0.95 for almost all cell lines, it is decided to employ ($\ell_1 \leq 6, \ell_2 = 2, \ell_3 \leq 6$) components for FE. Nonetheless, in the case of PC3 cells, ($\ell_1 \leq 8, \ell_2 = 2, \ell_3 \leq 8$), as an exception, are used for FE because the eighth component is found to have non-negligible contributions in this cell line.

To identify the genes and compounds associated with a significant dose-dependent activity, it is assumed that $u_{\ell_1 \leq 6, i}$ and $u_{\ell_3 \leq 6, k}$ follow independent normal distributions and P-values are attributed to the ith gene and the kth compounds using a χ^2 distribution,

$$P_i = P_{\chi^2}\left[> \sum_{\ell_1 \leq 6} \left(\frac{u_{\ell_1 i}^{(i)}}{\sigma_{\ell_1}} \right)^2 \right] \tag{7.5}$$

and

$$P_k = P_{\chi^2}\left[> \sum_{\ell_3 \leq 6} \left(\frac{u_{\ell_3 k}^{(k)}}{\sigma_{\ell_3}} \right)^2 \right] \tag{7.6}$$

where σ_{ℓ_1} and σ_{ℓ_3} are the standard deviations of $u_{\ell_1 i}^{(i)}$ and $u_{\ell_3 k}^{(k)}$, respectively. For PC3 cells, $\ell_1 \leq 8$ and $\ell_3 \leq 8$ are used in the above equations. $P_{\chi^2}[> x]$ is the cumulative probability that the argument is greater than x assuming a χ^2 distribution with eight degrees of freedom for PC3 cell lines and with six degrees of freedom for other cell lines. P_i and P_k are adjusted by means of the BH criterion, and compounds and genes associated with the adjusted P-value lower than 0.01 are selected as those associated with a significant dose-dependent cellular response. The number of selected genes and compounds are listed as "inferred genes" and "inferred compounds" in Table 7.8, respectively. The above process is illustrated in Fig. 7.4.

The next task is to identify proteins to which selected compounds bind. "inferred genes" in Table 7.8 do not correspond to the proteins to which selected compounds bind, because they are the genes whose mRNA expression is altered because of drug treatment. Usually, mRNA expression of proteins to which selected compounds bind is not altered because of drug treatment. Thus we need to infer proteins targeted by drug treatment. In order that, we need additional external information that lists the genes whose mRNA expression is altered because of a gene perturbation. Then if "inferred genes" matched with genes mRNA expression is altered because of the gene perturbation, we infer the perturbed gene as target protein (Fig. 7.5).

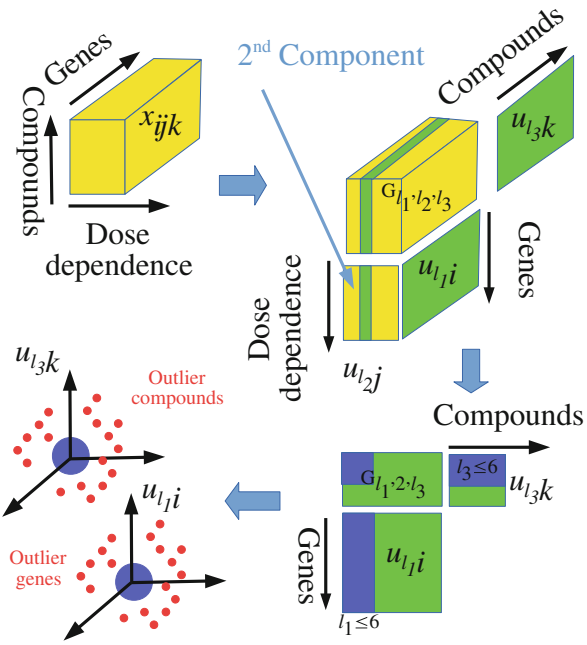

Fig. 7.4 Starting from gene expression profile formatted as tensor, x_{ijk}, singular value vectors, $\boldsymbol{u}_{\ell_1}^{(i)}$, $\boldsymbol{u}_{\ell_2}^{(j)}$, and $\boldsymbol{u}_{\ell_3}^{(k)}$, are obtained. After identifying $\ell_2 = 2$ as associated with linear dose dependence (see Fig. 7.3), $\ell_1 \leq 6$ and $\ell_3 \leq 6$ are decided to be used for FE because of larger contribution defined in Eq. (7.4). Genes i and compounds k are selected using $\boldsymbol{u}_{\ell_1}^{(i)}, \ell_1 \leq 6, \boldsymbol{u}_{\ell_3}^{(k)}, \ell_3 \leq 6$

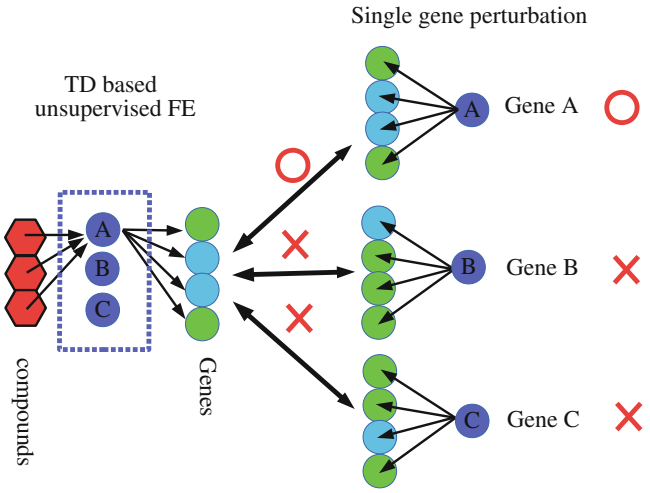

Fig. 7.5 After the drug (red hexagon) treatments, we can detect mRNAs with altered expression (filled cyan circle) along with those without altered expression (filled green circle). We have no information about proteins (circled A, B, and C). List of genes with altered expression can be compared with genes with altered expression when genes A, B, or C is perturbed. Then, we can identify compounds that might bind to protein A, because the list of genes whose mRNA expression is altered are common

There can be multiple resources from which we can retrieve the list of genes whose mRNA expression is altered because of single gene perturbation. Here we employ Enrichr [11] that collects multiple data resources in order to perform various enrichment analyses. After uploading "inferred genes" to Enrichr, we list genes associated with adjusted P-values less than 0.01 in the category of "Single gene Perturbations from GEO up." Their number corresponds to the number of "predicted targets" in Table 7.8. This strategy is especially efficient for LINCS data set that includes only expression of 978 genes. Employing the strategy in Fig. 7.5, we can identify target proteins not included in these 978 genes.

Next we would like to evaluate if our prediction is correct, i.e., if "inferred compounds" bind to "predicted targets." In principle, it is impossible to check the accuracy of our prediction without experiments. Thus, instead of executing experiments, we compare our prediction with known list of target proteins of drug compounds. For this purpose, we employ two information resources, drug2gene.com [19] and DSigDB [33]. Table 7.9 shows the results of Fisher's exact test that evaluates overlaps between "predicted targets" and known target proteins of "inferred compounds." If P-values computed by Fisher's exact test is less than 0.05, it is significant (no correction considering multiple comparisons). It is obvious that in most of the cases, our prediction significantly overlaps with known target proteins of drug compounds. Thus, TD based unsupervised FE can be used for in silico drug discovery from gene expression.

It is also interesting that "inferred compounds" are largely overlapped among cell lines. Because two to nine compounds are identified in each of 13 cell lines, the total number of identified compounds can be several tens. Nevertheless, the number of compounds listed in Table 7.9 is as small as 19. In some sense, it might be an evidence that our strategy is correct. It is reasonable that anti-cancer drugs are effective to multiple cancers. Thus, large overlap of "inferred compounds" between distinct cell lines makes sense. On the other hand, analyses based upon distinct gene expression profiles unlikely results in largely overlapped results without any biological reasons. Possibly, the result shown in Table 7.9 are trustable.

Although we employed single gene perturbation to infer target proteins from the list of genes with altered expression caused by drug treatment, any other database that can describe gene interaction should be usable. As an alternative, we try "PPI Hub Proteins" in Enrichr instead of "Single gene Perturbations from GEO up." The primary difference between "PPI Hub Proteins" and "Single gene Perturbations from GEO up" is the number of genes included. "PPI Hub Proteins" includes only a few hundred genes, while "Single gene Perturbations from GEO up" includes a few thousand genes. This suggests that the results using "PPI Hub Proteins" might be less significant. Table 7.10 lists the results of Fisher's exact test of the comparison between predicted targets based upon "PPI Hub Proteins" and drug2gene.com database. In contrast to the expectation, all cases have significant overlap with drug2gene.com. This supports our expectation that any kind of gene–gene interaction is usable together with TD based unsupervised FE for *in silico* drug discovery from gene expression.

Table 7.9 Compound–gene interactions presented in Table 7.8 that significantly overlap with interactions described in two data sets

Compounds	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Dabrafenib													○ ○
Dinaciclib							○ ○	○ ○	○ ○	○ ○	○ ○	○ ○	○ ○
CGP-60474			○ ×	○ ×	○ ×	○ ×	○ ×		○ ×			○ ×	○ ○
LDN-193189	○ ○				○ ○								○ ○
OTSSP167							– ○	– ○		– ○		– ○	– ○
WZ-3105		– ○		– ○	– ○		– ○	– ○	– ○			– ○	– ○
AT-7519				○ ○		○ ○		○ ○	○ ○			○ ○	
BMS-387032				○ ○		○ ○	○ ○		○ ○				
JNK-9L									○ ○				
Alvocidib	○ –	○ –	○ –	○ –	○ –	○ –			○ –				
GSK-2126458							– –					– –	
NVP-BEZ235							○ ×					○ ×	
Torin-2							× ○					× ○	
NVP-BGT226					– –			– –			– –	– –	
QL-XII-47	– –												
Celastrol	○ –												
A443654		○ ○		○ ○									
NVP-AUY922					× –	○ –							
Radicicol						○ –							

For each compound in the table, the upper row: the drug2gene.com data set is used for comparisons [19], the lower row: the DSigDB data set is used for comparisons [33]. Columns represent cell lines used in the analysis: (1) BT20, (2) HS578T, (3) MCF10A, (4) MCF7, (5) MDAMB231, (6) SKBR3, (7) A549, (8) HCC515, (9) HA1E,(10) HEPG2, (11) HT29, (12) PC3, (13) A375

○: a significant overlap between the data sets ($P < 0.05$); ×: no significant overlap between the data sets; –: no data; blank: no significant dose–response relation is identified. The confusion matrix and a full list of genes chosen in common are available in Additional file 3 [27].

Table 7.10 A significant overlap demonstrated between compound–target interactions presented in Table 7.8 and drug2gene.com

Compounds	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
										○	○	○	
CGP-60474			○	○	○	○	○		○			○	○
LDN-193189					○								
AT-7519				○		○		○	○			○	
BMS-387032				○		○	○		○				
Alvocidib		○	○	○	○	○			○				
NVP-BEZ235												○	
Celastrol	○												
A443654		○		○									
NVP-AUY922					○	○							
Radicicol						○							

In this case, the "PPI Hub Proteins" category in Enrichr is used. Labels (1) to (13) represent the same cell lines as described in Table 7.9

The full list of confusion matrices and genes chosen in common is available in Additional file 3 [27]

It might be useful to demonstrate how more direct and simple approach fails. One possible alternative simpler way is to apply linear regression

$$x_{ijk} = a_{ik} + b_{ik} D_j \tag{7.7}$$

where D_j is the jth dose density and a_{ik} and b_{ik} are regression coefficients. Then simply select i and k associated with more significant P-values as in the case of TD based unsupervised FE. In order to show that it cannot give us the reasonable set of is and ks, we apply Eq. (7.7) to A375 cell lines ((13) in Tables 7.8, 7.9, and 7.10) as an example. After correcting P-values that Eq. (7.7) gives by BH criterion, we find that all compounds have adjusted P-values less than 0.01 with at least one of the genes while all genes have adjusted P-values less than 0.01 with at least one of the compounds. Thus, by simply requesting "adjusted P-values less than 0.01" as in the case of TD based unsupervised FE, we cannot screen either genes or compounds. We can still try to select "top ranked" genes or compounds. In order to show that this cannot work well either, we apply two distinct criteria to select "top ranked" compounds as

- Select top ranked 10 compounds having larger number of genes associated with adjusted P-values less than 0.01.
- Suppose P_{ik} is P-value that Eq. (7.7) gives. Select top ranked 10 compounds having smaller $\sum_i \log P_{ik}$.

These two criteria rank compounds with more significant correlation with genes through dose density in some sense. The result is a bit disappointing (Table 7.11). Only three of top 10 compounds are chosen in common. This suggests that it is not

Table 7.11 Compounds selected by P-values that Eq. (7.7) gives, for A375 cell line ((13) in Tables 7.8, 7.9, and 7.10)

Compounds selected
Criterion 1
chelerythrine chloride, TGX-221, lapatinib, AS-601245, PIK-93, **canertinib**, LDN-193189, MK-2206, PF-04217903, **DCC-2036**
Criterion 2
ALW-II-49-7, AZ20, BI-2536, **canertinib**, celastrol, **chelerythrine chloride**, CHIR-99021, **DCC-2036**, dovitinib, GSK-1904529A

Bold ones are chosen in common

easy to select compounds in robust way simply based upon P-values that Eq. (7.7) gives. Thus, TD based unsupervised FE is much better strategy without no additional criterion than adjusted P-values than selection based upon P-values that Eq. (7.7) gives.

Before ending this section, I would like to mention briefly why the results of TD based unsupervised FE differ from that based upon linear regression, Eq. (7.7), so much in spite of that both TD based unsupervised FE and linear regression try to find the combinations of genes and compounds associated with dose dependence. As can be seen in Fig. 7.3, $\boldsymbol{u}_2^{(j)}$ used for FE is not simple linear function of dose density. In spite of that, the dependence of $\boldsymbol{u}_2^{(j)}$ upon dose density is quite universal, in other words, independent of cell lines. TD is the only method that can successfully identify this universal (independent of cell lines) functional form. There are no other ways to find it in advance. This cannot be achieved by any other supervised method, because any supervised method cannot avoid assuming something contradictory to this universal functional form. Because of this superiority, TD based unsupervised FE can achieve good performance shown in Tables 7.9 and 7.10.

7.4 Universarity of miRNA Transfection

miRNA transfection is a popular method that finds miRNA target genes experimentally. Nevertheless, some doubt arises if transfected miRNA can work similar to endogenous miRNAs [9], because it causes various unexpected effects that cannot be seen by upregulation of endogenous miRNAs. Because the aim of miRNA transfection experiments is to find miRNA target genes, only genes downregulated by the transfection are searched. Nevertheless, it is quite usual to find that many mRNAs are upregulated because of transfection. These upregulated mRNAs are usually ignored, because it is not interpretable from the knowledge about conventional miRNA functions. On the other hand, Jin et al. [9] argued that miRNA transfection can cause non-specific changes in gene expression. To the best of my knowledge, there are no studies that try to identify these non-specific effects in more positive points of view.

In this section, using TD based unsupervised FE, we are aiming to study how universal these non-specific gene expression alterations by miRNA transfections are. In order that, we collect multiple studies where multiple miRNA transfection experiments are performed. In individual studies, genes whose expression is altered in common over multiple miRNA transfection experiments are tried to be identified. Then it is checked if genes identified in individual studies are common over multiple studies. If so, sequence-nonspecific off-target regulation of mRNA does really exist and might play some critical roles in biology, too.

The identification of genes altered in common by sequence-nonspecific off-target regulation caused by miRNA transfection can be performed by TD based unsupervised FE as follows [26]. In usual application of TD based unsupervised FE, singular value vectors associated with desired sample dependence, e.g., distinction between patients and healthy controls, are searched to identify genes associated with such a dependence. On the contrary, in the present application, we are aiming to seek singular value vectors "not" associated with the distinction between transfected miRNAs, because lack of transfected miRNA dependence might be the evidence that gene expression alteration caused by miRNA expression toward these genes is because of sequence-nonspecific off-target regulation, no matter what the biological reasons that cause it are. Table 7.12 lists 11 studies including the gene expression profiles collected for the analysis in this study. It is obvious that they are quite diverse. Not only used cell lines but also transfected miRNAs differ from

Table 7.12 Eleven studies conducted for this analysis

Exp	GEO ID	Cell lines (cancer)	miRNA	Misc	Methods
1	GSE26996	BT549 (breast cancer)	miR-200a/b/c		PCA
2	GSE27431	HEY (ovarian cancer)	miR-7/128	mas5	PCA
3	GSE27431	HEY (ovarian cancer)	miR-7/128	plier	PCA
4	GSE8501	Hela (cervical cancer)	miR-7/9/122a/128a/132/133a /142/148b/181a		TD
5	GSE41539	CD1 mice	cel-miR-67, hsa-miR-590-3p, hsa-miR-199a-3p		PCA
6	GSE93290	multiple	miR-10a-5p, 150-3p/5p, 148a-3p/5p, 499a-5p, 455-3p		TD
7	GSE66498	multiple	miR-205/29a/144-3p/5p, 210,23b,221/222/223		TD
8	GSE17759	EOC 13.31 microglia cells	miR-146a/b	(KO/OE)	TD
9	GSE37729	HeLa	miR-107/181b	(KO/OE)	TD
10	GSE37729	HEK-293	miR-107/181b	(KO/OE)	TD
11	GSE37729	SH-SY5Y	181b	(KO/OE)	TD

More detailed information on how to process individual experiments in these eleven studies is available in Appendix. Methods: PCA or TD based unsupervised FE is used

experiments to experiments. Both KO (knock out) and OE (over expression) are considered. Thus, if there are genes chosen in common among these eleven studies, it is quite likely caused by sequence-nonspecific off-target regulation.

Because of their diversity, not only TD based unsupervised FE but also PCA based unsupervised FE is used. If the number of samples used for individual transfection in individual experiments does not match with one another, multiple experiments in which distinct miRNAs are transfected are hardly formulated in tensor forms. In these cases, PCA based unsupervised FE is employed instead. In the following, individual data set and how to format them in either matrix or tensor is discussed in a little bit detail in Appendix.

Table 7.13 shows the results. In spite of the heterogeneous data sets analyzed, they are highly consistent with one another. Thus, there might be some universal mechanisms that cause sequence-nonspecific off-target regulation.

From the data science point of view, it is important to see if other methods can derive the set of genes associated with the same amount of consistency among 11 studies listed in Table 6.12. For the comparison, we select t test. What we aim is essentially to find genes expressed distinctly between control and transfected samples. This kind of two class comparisons can be done by t test, too. In order to see if t test is inferior to TD and PCA based unsupervised FE, t test is applied to 11 studies. In this analysis, samples in individual studies are divided into two classes: samples to which no miRNAs (or mock miRNA) were transfected and samples to which miRNAs were transfected. Two-sided t test is applied to individual 11 studies. Then, obtained P-values are adjusted by BH criterion. Then, probes associated with adjusted P-values less than 0.01 are selected (Table 7.14). The result is a little bit disappointing. For five out of 11 studies, t test cannot identify any differently expressed genes. On the other hand, the numbers of selected genes vary from 35 to 11,060, which is contrast to the range of number of genes selected by PCA or TD based unsupervised FE, $\sim 10^2$ (Table 7.13). These numbers are unlikely biologically trustable. This possibly shows the failure of methodology.

In order to further demonstrate the inferiority of t test to TD or PCA based unsupervised FE, we try to reproduce the results of PCA or TD based unsupervised FE in Table 7.13. Since the number of genes selected by t test is often 0 (Table 7.14), the same number of top ranked genes with smaller P-values as those in PCA or TD based unsupervised FE are selected in individual experiments based upon P-values computed by t test even though P-values are not significant. It is obvious that the selected genes by t test are less coincident with each other than the selected genes by PCA or TD based unsupervised FE (Table 7.13) because odds ratios are smaller and P-values are larger. Thus, also from the point of coincidence between 11 studies, t test is inferior to TD or PCA based unsupervised FE.

Although PCA or TD based unsupervised FE successfully identifies sets of genes highly coincident between heterogeneous eleven studies, if they are not biologically reasonable, they are useless. In order to see biological values of selected genes, we here show one evaluation, although many evaluations were performed in my published paper [26] (I am not willing to show all of them here, because it might be simply boring).

Table 7.13 Fisher's exact test for coincidence among 11 miRNA transfection studies for PCA or TD based unsupervised FE and *t* test

Exp.		1	2	3	4	5	6	7	8	9	10	11
	#	232	711	747	441	123	292	246	873	113	104	120
PCA or TD based unsupervised FE												
1	232		4.14e−19	6.59e−22	3.96e−41	4.12e−71	9.41e−70	2.90e−60	1.34e−17	1.15e−27	6.84e−26	2.66e−07
2	711	7.68		0.00	1.89e−18	4.93e−27	5.59e−20	2.69e−32	4.62e−13	9.23e−16	8.66e−12	1.37e−03
3	747	8.30	345.52		3.63e−20	7.96e−21	5.70e−12	1.82e−27	9.52e−12	1.18e−14	1.01e−12	3.90e−06
4	441	18.23	5.19	5.34		6.14e−41	1.01e−34	1.44e−69	4.61e−11	2.16e−30	4.09e−28	1.35e−10
5	123	53.86	9.04	7.27	17.48		2.9e−179	1.27e−63	6.24e−15	3.16e−25	2.37e−17	4.69e−09
6	292	61.50	8.15	5.52	17.71	204.39		3.53e−53	2.57e−15	6.65e−22	1.65e−12	5.60e−05
7	246	20.27	5.35	4.67	12.39	20.11	22.03		6.91e−42	1.77e−36	4.50e−31	2.78e−14
8	873	18.61	7.22	6.51	8.29	15.61	18.53	20.73		1.81e−07	1.37e−06	2.76e−02
9	113	39.34	9.87	8.77	25.98	32.44	34.90	21.94	16.02		3.7e−125	9.27e−18
10	104	40.29	8.22	8.27	26.64	23.34	20.86	21.56	15.18	517.87		6.82e−16
11	120	10.15	3.19	4.43	9.19	11.55	8.11	8.28	4.92	19.57	18.70	
t test												
1	232		4.96e−04	8.49e−01	2.59e−01	6.35e−01	1.00e+00	5.40e−01	1.00e+00	4.08e−01	6.45e−01	6.68e−01
2	711	2.56		6.40e−69	1.38e−02	1.25e−01	1.55e−01	9.36e−03	1.00e+00	1.00e+00	3.76e−01	1.00e+00
3	747	0.80	10.49		8.65e−01	5.28e−01	3.76e−01	2.47e−01	7.79e−01	7.75e−01	5.30e−01	1.00e+00
4	441	1.55	1.90	0.89		6.58e−01	1.00e+00	4.31e−01	1.26e−01	2.71e−01	2.56e−01	1.00e+00
5	123	0.00	0.00	0.36	1.39		1.13e−22	1.00e+00	3.86e−01	1.00e+00	1.00e+00	1.00e+00
6	292	0.77	1.83	0.32	0.72	27.05		3.71e−01	1.00e+00	1.00e+00	1.00e+00	1.00e+00
7	246	1.16	0.48	0.71	1.22	0.67	0.31		4.47e−01	1.83e−01	7.60e−02	2.04e−01
8	873	0.64	1.00	1.17	2.15	2.09	0.00	0.46		1.59e−01	4.54e−01	1.27e−03
9	113	0.00	0.81	0.60	0.00	0.00	0.00	0.25	2.91		1.18e−03	4.07e−01
10	104	0.00	0.32	0.35	1.75	0.00	0.00	0.00	1.68	5.56		6.37e−01
11	120	1.31	0.78	0.88	0.97	0.00	0.00	1.69	6.87	0.00	0.00	

Upper triangle: *P*-value, lower triangle: odds ratio. #: number of genes selected in individual studies. "Xe−Y" means that "$X \times 10^{-Y}$"

Table 7.14 The number of genes selected by *t* test

Studies	1	2	3	4	5	6	7	8	9	10	11
Samples	6:6	3:4	6:4	18:18	2:2	16:16	19:19	18:18	6:12	6:12	4:4
Selected genes	11,060	0	0	0	0	35	280	55	5949	5730	0

Two numbers besides colon are the number of control and transfected samples, respectively

Table 7.15 is the result for KEGG pathway enrichment by uploading selected genes to Enrichr. It is obvious that not only there are many significant enrichment but also they are highly coincident between 11 studies. Thus, coincidence of selected genes between eleven studies shown in Table 7.13 is also biologically reasonable. In this sense, PCA or TD based unsupervised FE can identify biologically meaningful genes chosen in common between heterogeneous studies including various miRNAs transfected to various cell lines. Universal nature detected has seemingly biological importance, too.

7.5 One-Class Differential Expression Analysis for Multiomics Data Set

In general, there are two kinds of biological experiments, in vivo and in vitro. In vivo means real biological experiments using living organisms, e.g., animals and plants. Nevertheless, in vivo cannot be said as very economical, because it wastes whole body even when we are interested in a specific tissue. For example, even if you are interested in liver disease, in vivo experiments require to cultivate a whole body. You may wonder if only liver can be separately cultivated, it would be more effective. In vivo experiments recently have tendency to be avoided from the ethical point of view, too, because they kill numerous animals. In vitro experiments can fulfill these requirements more or less. in vitro makes use of cell lines, which is an immortalized cell that is often made out of cancer cells. Once cell line is established, you can do any kind of experiments in vitro using cell lines. Because cell lines can be cultivated even in a dish, it is definitely cost effective and does not kill any animals.

One possible problem of in vitro is the lack of control samples. It is known that cell lines differ from the tissue cells from which cell lines are established. Thus, usually cell lines are compared between not treated and treated ones. Characterizing immortalized cell lines themselves is not an easy task.

In this section, we propose the method that can characterize cancer cell line from gene expression without comparing with something [22]. In this criterion, genes are expressive in common over multiple cancer subtypes are searched and are considered to be characteristic gene expression of cancer cell line. In this regard, TD based unsupervised FE used to identify expressed gene in common over multiple miRNAs transfection studies in the previous section is employed again.

Table 7.15 In each of 11 studies, 20 top-ranked significant KEGG pathways whose associated genes significantly match some genes selected for each experiment are identified

Exp	#	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)
1	(232)	31/137	7/168	10/142	6/133	9/55	9/193			7/169	8/203
	[10]	3.69e−29	3.18e−02	1.66e−04	3.45e−02	1.02e−06	6.85e−03			3.18e−02	3.01e−02
2	(711)	36/137	18/168	14/142	12/133	13/55				16/169	18/203
	[12]	3.43e−19	1.48e−03	1.05e−02	3.20e−02	5.92e−06				8.12e−03	8.12e−03
3	(747)	23/137	15/168			14/55				18/169	19/203
	[15]	3.58e−07	1.94e−02			1.20e−06				2.02e−03	4.78e−03
4	(441)	50/137	15/168	19/142	18/133	6/55	19/193	7/78	12/151	9/169	
	[10]	2.92e−45	1.91e−04	3.97e−08	6.42e−08	2.49e−02	3.40e−06	2.74e−02	4.44e−03	1.29e−01	
5	(123)	9/137						8/78		6/169	8/203
	[23]	2.97e−06						6.08e−07		4.29e−03	3.03e−04
6	(292)	45/137	20/168	19/142	18/133	4/55	19/193	11/78	12/151		
	[14]	1.35e−46	3.32e−11	2.27e−11	4.00e−11	7.95e−02	2.24e−09	4.90e−07	4.87e−05		

7	(246)	40/137	9/168	10/142	9/133		11/193	4/78	7/151		6/203
	[7]	5.61e−42	6.60e−03	5.80e−04	1.32e−03		1.16e−03	2.57e−01	6.31e−02		4.52e−01
8	(873)	75/137	30/168	32/142	32/133		36/193	14/78	24/151	25/169	
	[24]	5.59e−63	2.09e−09	9.32e−13	1.89e−13		7.51e−12	1.39e−04	1.62e−06	3.11e−06	
9	(113)	18/137	11/168	12/142	10/133	6/55	12/193	4/78	11/151		
	[20]	8.24e−18	7.10e−08	1.66e−09	8.42e−08	8.85e−06	2.96e−08	6.64e−03	2.96e−08		
10	(104)	11/137	8/168	9/142	8/133	5/55	10/193		8/151		
	[20]	1.98e−08	6.68e−05	3.23e−06	1.71e−05	1.56e−04	3.23e−06		3.60e−05		
11	(120)	6/137		4/142		5/55					5/203
	[3]	9.04e−03		8.49e−02		2.98e−03					6.83e−02

Thus, the following KEGG pathways are most frequently ranked within the top 20. "Xe−Y" means that "$X \times 10^{-Y}$"
(i) Ribosome:hsa03010, (ii) Alzheimer's disease:hsa05010, (iii) Parkinson's disease:hsa05012, (iv) Oxidative phosphorylation:hsa00190, (v) Pathogenic *Escherichia coli* infection:hsa05130, (vi) Huntington's disease:hsa05016, (vii) Cardiac muscle contraction:hsa04260, (viii) Nonalcoholic fatty liver disease (NAFLD):hsa04932, (ix) Protein processing in endoplasmic reticulum:hsa04141, and (x) Proteoglycans in cancer:hsa05205. (numbers):gene, [numbers]:KEGG pathways associated with adjusted P-values less than 0.01. Upper rows in each exp: (the number of genes coinciding with the genes selected for each experiment)/(genes listed in Enrichr in each category). Lower rows in each exp: adjusted P-values provided by Enrichr

In addition to this, TD based unsupervised FE is used as a tool that integrates omics data. The data set used is downloaded from DBTSS [21], which is a database of transcriptional start sites (TSS), and includes RNA-seq, TSS-seq, and ChIP-seq (histone modification, H3K27ac). These are observed in 26 NSCLC subtype cell lines using HTS technology; DBTSS also stores various omics data set measured on various cell lines and living organisms.

Before starting analysis, we briefly explain the difference among TSS-seq, RNA-seq, and ChIP-seq. As it name says, TSS-seq tries to sequence RNA transcribed from the region around TSS. Thus, TSS-seq basically counts how many times transcription starts. On the other hand, RNA-seq counts the fragments taken from any part of whole RNA. In this sense, RNA-seq counts the total amount of RNA transcribed. Generally, TSS-seq and RNA-seq are positively correlated, although there are no known functional forms that relate between these two, because the function is affected by many factors, e.g., individual genes have various length and some genes are long while others are short. If longer genes are more transcribed, the ratio RNA-seq to TSS-seq becomes larger. In addition to this, individual genes have isoforms, each of which has different length. This mechanism is called as an alternative splicing. If more number of longer isoforms are transcribed from each gene, it also contributes to the increased RNA-seq/TSS-seq ratio. Although there are many detailed points that must be considered in order to relate RNA-seq to TSS-seq, there is one clear point; TSS-seq and RNA-seq should be positively correlated. Thus, seeking genes associated with both more TSS-seq counts and RNA-seq counts can reduce the possibility that genes are wrongly identified as being upregulated or downregulated, e.g., because of technical issues like miss amplification.

ChIP-seq is a different technology that detects to which part of DNA the protein binds. Although I do not explain the details of the relationship between DNA and proteins that bind to it, basically DNA binding protein can control the rate of transcription. ChIP-seq can study this relationship by considering DNA binding protein. Histone modification is more advanced feature. In order to suppress the self-entanglements of lengthy DNA, long DNA string is wrapped around protein core called histone. Because tightly wrapped DNA is hardly transcribed, how tightly DNA is wrapped around histone can affect the amount of transcription drastically. On the other hand, affinity between histone and DNA can be affected by chemical modification of histone. Among various histone modification, acetylation of histone tail is supposed to enhance the transcription by reducing the affinity between DNA and histone. As a result, considering histone modification (H3K27ac) together with RNA-seq and TSS-seq can further reduce the possibility of wrongly identified up/downregulated genes. In the following, we try to seek genes simultaneously associated with the increased TSS-seq, RNA-seq, and ChIP-seq that measureds H3K27ac counts.

When formatting RNA-seq, TSS-seq, and ChIP-seq measurement data into tensor form, how we can practically perform this is a problem. Fundamentally, although it is possible to perform it in single nucleotide base, it results in too huge tensor that requires too large memory to manage. In this case, it is better to employ coarse graining approach that takes average over local chromosome regions. The

problem is how long regions should be. If the length of the region is too large, each region includes more than one (protein coding) genes. Then, increased or decreased counts within each region might reflect more than one genes. This will result in low interpretability. On the other hand, if the length of the region is too short, individual (protein coding) genes are expressed over multiple region. It again results in low interpretability. Thus, there should be somewhat optimal length of region. In this section, I try 25,000 nucleotides as a length of region. Generally, the average length of protein is $\sim 10^2$. Because one amino acid is coded by three-nucleotide (codon), a length of region that codes individual protein coding genes should be at most $\sim 10^3$. The regions that code protein coding genes are typically composed of both exon and intron, which correspond to translated and non-translated regions, respectively. Thus, the region of DNA that codes individual genes might be doubled. It is still expected not to exceed $\sim 10^3$ so much. In actuality, some literature reported that average length of DNA regions that code human protein coding genes is still a little bit shorter than $\sim 10^4$ [8]. Nevertheless, if the region over which TSS-seq, RNA-seq and ChiP-seq count data is averaged is as long as expected length of DNA region that codes individual protein coding genes, boundaries between averaging region might frequently fall into the mid of the DNA region that codes individual protein coding region. Thus, the length of region averaging counts data should be a few times longer than expected length of DNA region that codes individual protein coding region. Based upon these considerations, 25,000 nucleotides region over which TSS-seq, RNA-seq, and ChIP-seq counts are averaged is proposed.

In the data set having a type "human lung adenocarcinoma cell line 26 cell line" in inhouse data category, RNA-seq, TSS-seq, and ChIP-seq data are used. Among ChIP-seq data, only the H3K27ac is used (H3K27ac means that K27 position of the 3rd histone (H3) is acetlyated). Counts are averaged over chromosomal regions fragmented to regions of length of 25,000 nucleotides. Tensors are generated for each chromosome separately. Then, tensor is the form of $x_{ijk} \in \mathbb{R}^{N \times 26 \times 3}$, where N is the total number of regions of the length of 25,000 nucleotides within each chromosome, j stands for 26 cell lines, and k stands for counts of TSS-seq, RNA-seq, and ChIP-seq. HOSVD algorithm, Fig. 3.8, is applied to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{26} \sum_{\ell_3=1}^{3} G(\ell_1, \ell_2, \ell_3) u^{(i)}_{\ell_1 i} u^{(j)}_{\ell_2 j} u^{(k)}_{\ell_3 k} \tag{7.8}$$

where $u^{(i)}_{\ell_1 i} \in \mathbb{R}^{N \times N}$, $u^{(j)}_{\ell_2 j} \in \mathbb{R}^{26 \times 26}$, and $u^{(k)}_{\ell_3 k} \in \mathbb{R}^{3 \times 3}$ are singular value matrices and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{N \times 26 \times 3}$ is a core tensor.

First, we need to find $\boldsymbol{u}^{(j)}_{\ell_2}$ that is independent of 26 cell lines and $\boldsymbol{u}^{(k)}_{\ell_3}$ that is independent of RNA-seq, TSS-seq, and ChIP-seq. Figure 7.6 shows $\boldsymbol{u}^{(j)}_1$. Excluding X chromosome, it is highly independent of 26 cell lines. Then we decide to employ $\ell_2 = 1$. Figure 7.7 shows $\boldsymbol{u}^{(k)}_1$. They are highly independent of TSS-seq, RNA-seq, and ChIP-seq. Then we decide to employ $\ell_3 = 1$.

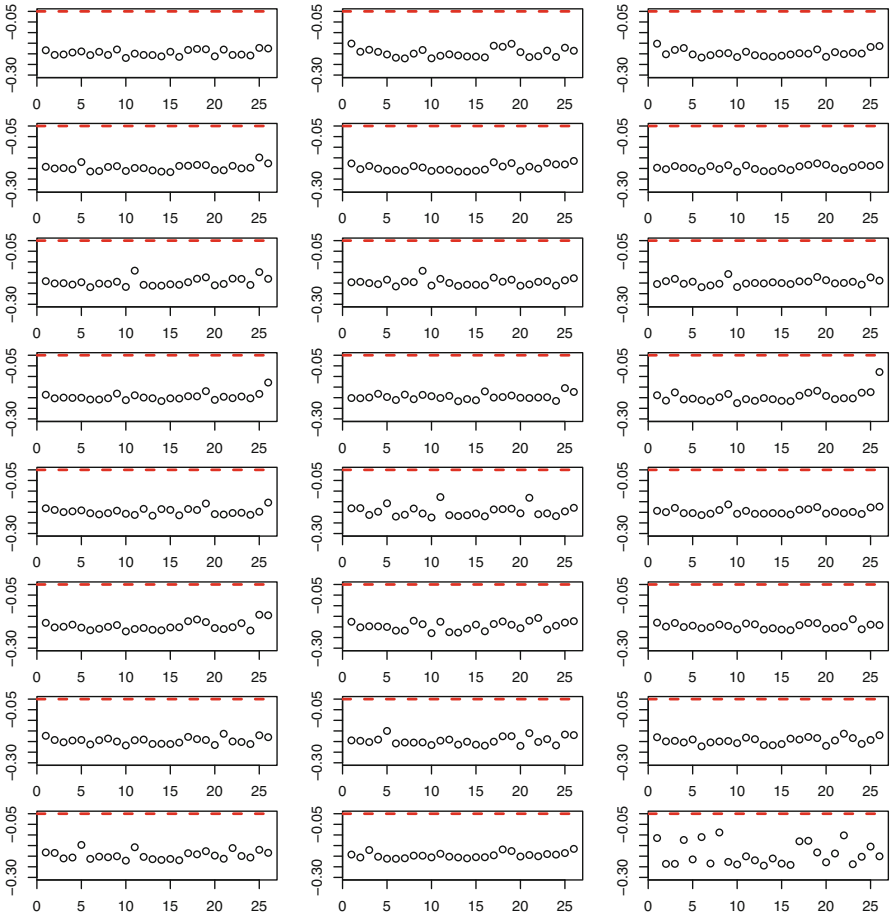

Fig. 7.6 $\boldsymbol{u}_1^{(j)}$. The first row, from left to right, chromosome 1, 2, 3, the second row, from left to right, chromosome 4, 5, 6 , and so on. The last row, from left to right, chromosome 22, X, Y. Red broken line is baseline

Then we try to find which $G(\ell_1, 1, 1)$ has the largest absolute value and find that $G(1, 1, 1)$ has always the largest absolute values independent of chromosome. Thus, $\boldsymbol{u}_1^{(i)}$ is used to attributed P-value to regions as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{1i}^{(i)}}{\sigma_1} \right)^2 \right]. \tag{7.9}$$

P-values are collected from 24 chromosome and are corrected by BH criterion. Then 826 regions associated with adjusted P-values less than 0.01 are selected. 826 is very small compared with the total number of regions; because the total number

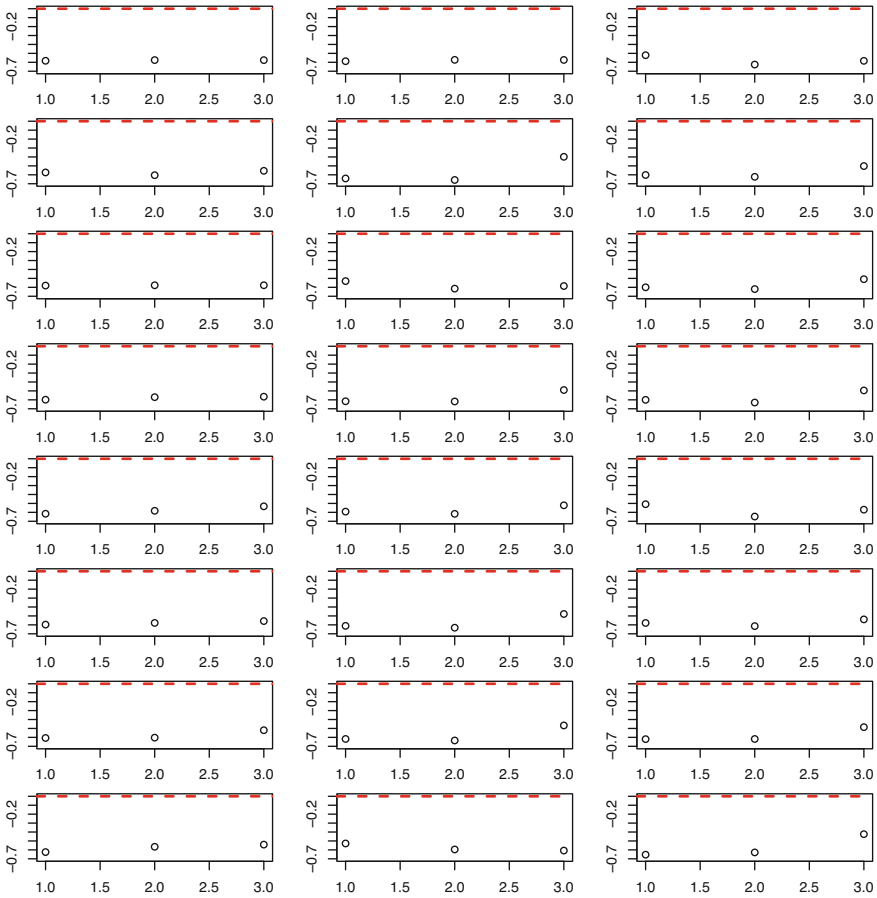

Fig. 7.7 $\boldsymbol{u}_1^{(k)}$. The first row, from left to right, chromosome 1, 2, 3, the second row, from left to right, chromosome 4, 5, 6, and so on. The last row, from left to right, chromosome 22, X, Y. Red broken line is baseline

of regions is about $3 \times 10^9/2.5 \times 10^4 \sim 10^5$ where 3×10^9 is the total length of human genome while 2.5×10^4 is the length of individual regions, 826 corresponds to as little as 0.8% of regions. This is reasonable because only a few percentages of genome code protein coding genes.

In order to validate these selected regions, we upload 1741 Entrez genes associated with these 826 regions to DAVID. Entrez genes are gene ID manually curated gene unique ID that is integer number [12]. Table 7.16 lists the KEGG pathway enrichment associated with adjusted P-values less than 0.05. At a glance, they do not look like related to cancers. Nevertheless, some of them are cancer related terms. For example, the relationship between "antigen processing and presentation" and cancer is often discussed [4]. Parkinson's disease is often reported to be related

Table 7.16 KEGG pathway enrichment by the 1741 Entrez genes identified by TD based unsupervised FE

Category	Term	Genes count	%	P-value	Adjusted P-value
KEGG_PATHWAY	Ribosome	73	4.2	9.8×10^{-38}	2.7×10^{-35}
KEGG_PATHWAY	Spliceosome	39	2.2	6.2×10^{-10}	8.4×10^{-8}
KEGG_PATHWAY	Protein processing in endoplasmic reticulum	41	2.4	8.0×10^{-8}	7.3×10^{-6}
KEGG_PATHWAY	Antigen processing and presentation	22	1.3	8.0×10^{-6}	5.5×10^{-4}
KEGG_PATHWAY	Pathogenic Escherichia coli infection	17	1.0	1.7×10^{-5}	9.2×10^{-4}
KEGG_PATHWAY	Parkinson's disease	30	1.7	9.6×10^{-5}	4.3×10^{-3}
KEGG_PATHWAY	Biosynthesis of antibiotics	39	2.2	1.6×10^{-4}	6.3×10^{-3}
KEGG_PATHWAY	Oxidative phosphorylation	26	1.5	1.0×10^{-3}	3.5×10^{-2}
KEGG_PATHWAY	Bacterial invasion of epithelial cells	18	1.0	1.2×10^{-3}	3.6×10^{-2}
KEGG_PATHWAY	Alzheimer's disease	30	1.7	1.7×10^{-3}	4.6×10^{-2}

Adjusted P-values are by BH criterion

to lung cancer [30]. Although we are not willing to discuss fully about the relations between the detected KEGG pathway enrichment and NSCLC, it is obvious that TD based unsupervised FE can detect set of genes including those related to NSCLC.

Although it is better to evaluate the performance of TD based unsupervised FE based upon the comparison with other methods, it is not easy because there are no control samples to be compared. Thus, alternatively we select genes based upon the ratio of standard deviation to average over 26 cell lines, because the smaller ratio of variance to mean might suggest smaller variability between 26 cell lines. For each of TSS-seq, RNA-seq, and ChIP-seq, we select top 5% regions with smaller ratio. Then regions chosen in common among TSS-seq, RNA-seq, and ChIP-seq are collected; we find that 2041 Entrez genes are included in these regions chosen in common. This number, 2041, is comparative with 1741 that is the number of Entrez genes selected by TD based unsupervised FE. Thus, uploading these to DAVID is a suitable test to see if TD based unsupervised FE is superior to this alternative method. Then we find that only two KEGG pathways, "Spliceosome" and "Ubiquitin mediated proteolysis" are associated with adjusted P-values less than 0.05. This suggests that TD based unsupervised FE can identify far more biologically reasonable set of genes than this alternative approach.

7.6 General Examples of Case I and II Tensors

Before demonstrating individual cases using case I and case II tensor in detail, we demonstrate various cases briefly based upon the recent publication [23]. As shown in Table 5.3, matrices or low mode tensor can be combined to generate (higher mode) tensor. In this section, we demonstrate how the combinations shown in Table 5.3 work to select genes critical to the diseases or phenomena considered.

7.6.1 Integrated Analysis of mRNA and miRNA

Integrated analysis of mRNA and miRNA was also performed by PCA based unsupervised FE (Sect. 6.4), which is once applied to mRNA and miRNA separately. Then obtained two sets of PC loading attributed to sample were investigated to seek those sharing common nature between two sets. After that, corresponding PC scores attributed to mRNA and miRNA were used for FE. On the contrary, in the application of TD based unsupervised FE to the integrated analysis of mRNA and miRNA, mRNA and miRNA expression profiles are integrated in advance.

The analyzed data set is composed of mRNA and miRNA profiles which were measured for multi-class breast cancer samples including normal breast tissues [7]. mRNA and miRNA expression profiles of multi-omics data are downloaded from GEO using GEO ID GSE28884. At first, GSE28884_RAW.tar is downloaded and expanded. For mRNA, 161 files whose names ended by the string "c.txt.gz" are used. Each file is loaded into R by read.csv command and the second column named "M" is employed as mRNA expression values. Probes not associated with Human Genome Organisation (HUGO) gene names are discarded and 13,393 probes remain. One hundred and sixty one files whose names end by the string "geo.txt.gz" are used for miRNA expression profiles; mRNA expression profiles of the corresponding samples are also used. Each file is loaded into R by read.csv command and the second column ("Count") is summed using the same third column ("Annotation") values. If the resulting total sum is less than 10, it is discarded and not used for further analysis.

Because the 161 samples are shared between miRNA and mRNA expression profiles, the multi-omics data corresponds to case I data (Table 5.3). TD based unsupervised FE is applied to the data set in order to identify disease critical genes and latent relations between miRNA and mRNA, whose expression profiles are $x_{i_1 j}^{\mathrm{mRNA}} \in \mathbb{R}^{13393\times 161}$ and $x_{i_2 j}^{\mathrm{miRNA}} \in \mathbb{R}^{755\times 161}$, respectively. They can be formatted as case I tensor as

$$x_{i_1 i_2 j} = x_{i_1 j}^{\mathrm{mRNA}} x_{i_2 j}^{\mathrm{miRNA}}. \tag{7.10}$$

HOSVD, Fig. 3.8, is applied to $x_{i_1 i_2 j}$ as

$$x_{i_1 i_2 j} = \sum_{\ell_1=1}^{13393} \sum_{\ell_2=1}^{755} \sum_{\ell_3=1}^{161} G(\ell_1, \ell_2, \ell_3) u^{(i_1)}_{\ell_1 i_1} u^{(i_2)}_{\ell_2 i_2} u^{(j)}_{\ell_3 j} \tag{7.11}$$

where $u^{(i_1)}_{\ell_1 i_1} \in \mathbb{R}^{13393\times 13393}$,$u^{(i_2)}_{\ell_2 i_2} \in \mathbb{R}^{755\times 755}$ and $u^{(j)}_{\ell_3 j} \in \mathbb{R}^{161\times 161}$ are singular value matrices and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{13393\times 755\times 161}$ is a core tensor.

First we need to seek singular value vectors, $\boldsymbol{u}^{(j)}_{\ell_3} \in \mathbb{R}^{161}$, with significant cancer subtype dependence. Figure 7.8 shows boxplots of $\boldsymbol{u}^{(j)}_{\ell_3}$, $1 \le \ell_3 \le 5$; it is obvious that these singular value vectors have significant class (cancer subtypes) dependence. The next step is to find $G(\ell_1, \ell_2, 1 \le \ell_3 \le 5)$ with larger absolute values. Table 7.17 shows the top ranked $G(\ell_1, \ell_2, 1 \le \ell_3 \le 5)$s; there are clearly only $1 \le \ell_1 \le 5$ and $1 \le \ell_2 \le 2$, respectively. Thus, P-values are attributed to i_1 and i_2 using $u^{(i_1)}_{\ell_1 i_1}$, $1 \le \ell_1 \le 5$ and $u^{(i_2)}_{\ell_2 i_2}$, $1 \le \ell_1 \le 2$, respectively, as

$$P_{i_1} = P_{\chi^2}\left[> \sum_{\ell_1=1}^{5} \left(\frac{u^{(i_1)}_{\ell_1 i_1}}{\sigma_{\ell_1}} \right)^2 \right], \tag{7.12}$$

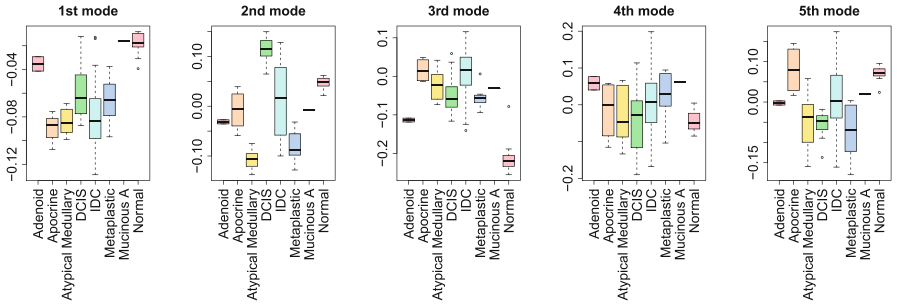

Fig. 7.8 Boxplot of $\boldsymbol{u}^{(j)}_{\ell_3}$, $1 \le \ell_3 \le 5$ when HOSVD is applied as Eq. (7.11). P-values computed by categorical regression. 1st: 2.39×10^{-5}, 2nd: 5.83×10^{-14}, 3rd: 1.36×10^{-24}, 4th: 2.58×10^{-2}, 5th: 2.12×10^{-5}

Table 7.17 Top ranked 10 $G(\ell_1, \ell_2, 1 \le \ell_3 \le 5)$s with larger absolute values among $1 \le \ell_1, \ell_2, \ell_3 \le 10$ in Eq. (7.11)

ℓ_1	1	2	4	3	5
ℓ_2	1	1	1	1	1
ℓ_3	1	2	4	3	5
$G(\ell_1, \ell_2, \ell_3)$	1.67×10^5	-1.03×10^5	7.48×10^4	-6.64×10^4	6.23×10^4
ℓ_1	3	1	3	2	1
ℓ_2	2	2	1	2	2
ℓ_3	3	3	5	3	2
$G(\ell_1, \ell_2, \ell_3)$	3.00×10^4	-2.87×10^4	-2.33×10^4	-2.02×10^4	-1.48×10^4

$$P_{i_2} = P_{\chi^2}\left[> \sum_{\ell_2=1}^{2} \left(\frac{u^{(i_2)}_{\ell_2 i_2}}{\sigma_{\ell_2}} \right)^2 \right]. \tag{7.13}$$

Computed P-values are adjusted by BH criterion; i_1s and i_2s associated with adjusted P-values less than 0.01 are selected. Then, 426 mRNA probes and 7 miRNAs are selected, respectively.

In order to evaluate selected 426 mRNAs biologically, we upload these mRNAs to DAVID. Then we can find numerous enrichment. Tables 7.18 and 7.19 show the results of GO term enrichment (adjusted P-values less than 0.05). BP is related to biological feature, CC is related to the location within cell, and MF is function of gene as molecules. Although we are not willing to summarize all of them, most of them are reasonably related to cancers, e.g., immune related or cell surface enrichment. Thus TD based unsupervised FE is likely successful to identify cancer related genes.

In order to demonstrate superiority of type I tensor, we also employ type II tensor as

$$x_{i_1 i_2} = \sum_j x_{i_1 i_2 j}. \tag{7.14}$$

Table 7.18 GO BP enrichment by the 426 ensembl genes identified by TD based unsupervised FE

Category	Term	Genes count	%	P-value	Adjusted P-value
GOTERM_BP_DIRECT	Immune response	36	11.4	2.7×10^{-14}	5.6×10^{-11}
GOTERM_BP_DIRECT	Signal transduction	57	18.1	5.1×10^{-12}	5.3×10^{-9}
GOTERM_BP_DIRECT	Type I interferon signaling pathway	10	3.2	1.8×10^{-6}	1.2×10^{-3}
GOTERM_BP_DIRECT	Collagen catabolic process	10	3.2	1.8×10^{-6}	1.2×10^{-3}
GOTERM_BP_DIRECT	Positive regulation of cell proliferation	25	7.9	3.2×10^{-6}	1.3×10^{-3}
GOTERM_BP_DIRECT	Cell–cell signaling	18	5.7	3.1×10^{-6}	1.6×10^{-3}
GOTERM_BP_DIRECT	Response to estradiol	11	3.5	4.8×10^{-6}	1.6×10^{-3}
GOTERM_BP_DIRECT	Defense response to virus	14	4.4	8.0×10^{-6}	2.3×10^{-3}
GOTERM_BP_DIRECT	B cell receptor signaling pathway	8	2.5	4.5×10^{-5}	1.1×10^{-2}
GOTERM_BP_DIRECT	Positive regulation of cAMP metabolic process	4	1.3	1.1×10^{-4}	2.4×10^{-2}
GOTERM_BP_DIRECT	Response to peptide hormone	7	2.2	1.2×10^{-4}	2.4×10^{-2}

(continued)

Table 7.18 (continued)

Category	Term	Genes count	%	P-value	Adjusted P-value
GOTERM_BP_DIRECT	Negative regulation of apoptotic process	21	6.7	1.9×10^{-4}	2.6×10^{-2}
GOTERM_BP_DIRECT	Defense response	8	2.5	1.8×10^{-4}	2.6×10^{-2}
GOTERM_BP_DIRECT	T cell activation	7	2.2	1.7×10^{-4}	2.7×10^{-2}
GOTERM_BP_DIRECT	T cell differentiation	6	1.9	1.7×10^{-4}	2.8×10^{-2}
GOTERM_BP_DIRECT	Skeletal system development	11	3.5	1.7×10^{-4}	3.1×10^{-2}
GOTERM_BP_DIRECT	Chemokine-mediated signaling pathway	8	2.5	2.6×10^{-4}	3.3×10^{-2}
GOTERM_BP_DIRECT	Mast cell activation	4	1.3	2.9×10^{-4}	3.4×10^{-2}
GOTERM_BP_DIRECT	Adaptive immune response	11	3.5	3.1×10^{-4}	3.5×10^{-2}
GOTERM_BP_DIRECT	Cell surface receptor signaling pathway	15	4.8	4.0×10^{-4}	4.2×10^{-2}
GOTERM_BP_DIRECT	Cellular response to interferon-alfa	4	1.3	4.3×10^{-4}	4.3×10^{-2}
GOTERM_BP_DIRECT	Inflammatory response	18	5.7	4.5×10^{-4}	4.3×10^{-2}
GOTERM_BP_DIRECT	Apoptotic process	23	7.3	5.2×10^{-4}	4.5×10^{-2}
GOTERM_BP_DIRECT	Humoral immune response	7	2.2	5.0×10^{-4}	4.6×10^{-2}
GOTERM_BP_DIRECT	Positive regulation of neutrophil chemotaxis	5	1.6	5.5×10^{-4}	4.6×10^{-2}
GOTERM_BP_DIRECT	Collagen fibril organization	6	1.9	6.0×10^{-4}	4.8×10^{-2}
GOTERM_BP_DIRECT	Proteolysis	21	6.7	6.4×10^{-4}	4.9×10^{-2}

Adjusted P-values are by BH criterion

Applying SVD to $x_{i_1 i_2}$, we get singular value vectors $u^{(i_1)}_{\ell_1 i_1} \in \mathbb{R}^{13393 \times 161}$ and $u^{(i_2)}_{\ell_2 i_2} \in \mathbb{R}^{755 \times 161}$. In order to select singular vector used for FE, we need to know dependence upon classes (in this case, cancer subtype). In order that, we need singular value vectors attributed to samples. It is computed as Eqs. (5.12) and (5.13),

$$u^{j;i_1}_{\ell_1 j} = \sum_{i_1=1}^{13393} x_{i_1 j} u^{(i_1)}_{\ell_1 i_1} \tag{7.15}$$

$$u^{j;i_2}_{\ell_2 j} = \sum_{i_2=1}^{755} x_{i_2 j} u^{(i_2)}_{\ell_2 i_2} \tag{7.16}$$

Figure 7.9 shows boxplot of $u^{j;i_1}_{\ell_1 j}$ and $u^{j;i_2}_{\ell_2 j}$ for $1 \le \ell_3 \le 5$. It is obvious that these singular value vectors have significant class (cancer subtypes) dependence.

Thus, P-values are attributed to i_1 and i_2 using $u^{(i_1)}_{\ell_1 i_1}$ and $u^{(i_2)}_{\ell_2 i_2}$ for $1 \le \ell_3 \le 5$, respectively, as

Table 7.19 GO CC and MF enrichment by the 426 ensembl genes identified by TD based unsupervised FE

Category	Term	Genes count	%	P-value	Adjusted P-value
GOTERM_CC_DIRECT	Extracellular space	84	26.7	1.60×10^{-26}	4.90×10^{-24}
GOTERM_CC_DIRECT	Extracellular region	82	26	3.10×10^{-20}	4.80×10^{-18}
GOTERM_CC_DIRECT	Extracellular exosome	97	30.8	9.00×10^{-13}	9.20×10^{-11}
GOTERM_CC_DIRECT	External side of plasma membrane	23	7.3	1.00×10^{-11}	7.70×10^{-10}
GOTERM_CC_DIRECT	Cell surface	23	7.3	1.20×10^{-4}	7.40×10^{-3}
GOTERM_CC_DIRECT	Extracellular matrix	15	4.8	4.80×10^{-4}	1.80×10^{-2}
GOTERM_CC_DIRECT	Multivesicular body	5	1.6	4.40×10^{-4}	1.90×10^{-2}
GOTERM_CC_DIRECT	Anchored component of membrane	9	2.9	6.20×10^{-4}	2.10×10^{-2}
GOTERM_CC_DIRECT	Cytosol	80	25.4	4.20×10^{-4}	2.10×10^{-2}
GOTERM_MF_DIRECT	Protein homodimerization activity	34	10.8	8.60×10^{-7}	4.90×10^{-4}
GOTERM_MF_DIRECT	RAGE receptor binding	5	1.6	2.90×10^{-5}	5.50×10^{-3}
GOTERM_MF_DIRECT	Chemokine activity	8	2.5	2.40×10^{-5}	6.70×10^{-3}
GOTERM_MF_DIRECT	CXCR3 chemokine receptor binding	4	1.3	5.40×10^{-5}	7.60×10^{-3}
GOTERM_MF_DIRECT	Receptor binding	18	5.7	2.00×10^{-4}	1.90×10^{-2}
GOTERM_MF_DIRECT	Serine-type endopeptidase activity	15	4.8	2.00×10^{-4}	2.20×10^{-2}
GOTERM_MF_DIRECT	Protein binding	187	59.4	2.90×10^{-4}	2.30×10^{-2}
GOTERM_MF_DIRECT	Identical protein binding	28	8.9	4.00×10^{-4}	2.80×10^{-2}

Adjusted P-values are by BH criterion

$$P_{i_1} = P_{\chi^2}\left[> \sum_{\ell_1=1}^{5} \left(\frac{u_{\ell_1 i_1}^{(i_1)}}{\sigma_{\ell_1}} \right)^2 \right], \tag{7.17}$$

$$P_{i_2} = P_{\chi^2}\left[> \sum_{\ell_2=1}^{5} \left(\frac{u_{\ell_2 i_2}^{(i_2)}}{\sigma_{\ell_2}} \right)^2 \right]. \tag{7.18}$$

P-values are adjusted by BH criterion. i_1 and i_2 associated with adjusted P-values less than 0.01 are selected. Then, 374 mRNA probes and 21 miRNAs are selected.

In order to validate selected 374 mRNAs, we upload these mRNAs to DAVID. Then we can find numerous enrichment. Table 7.20 shows the results of GO term enrichment (adjusted P-values less than 0.05) as in Tables 7.18 and 7.19. Thus, although the number of enrichment decreases than that in the type I tensor, still there are many cancer related GO terms. Thus, type II tensor approach is still valid enough biologically.

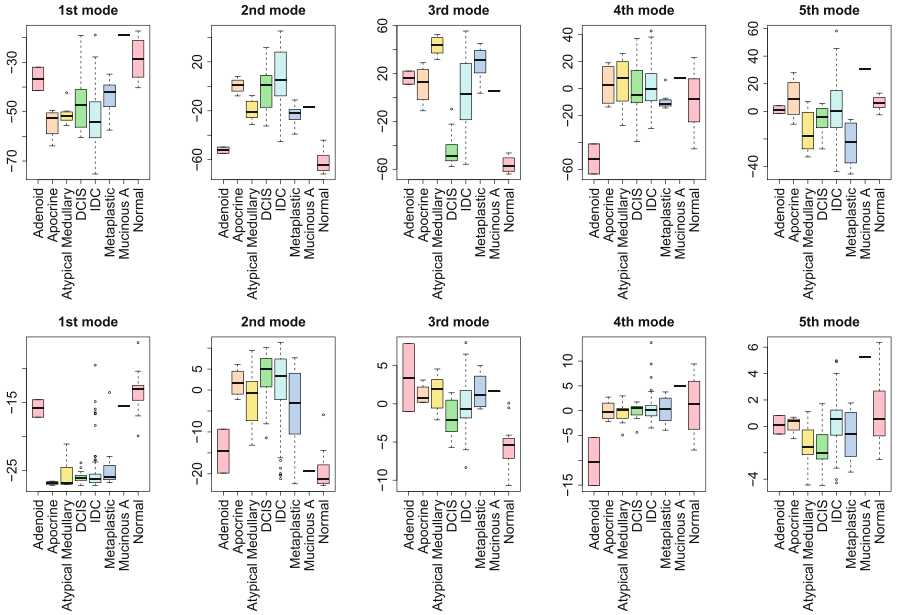

Fig. 7.9 Boxplot of $u^{j;i_1}_{\ell_1 j}$ (upper row) and $u^{j;i_2}_{\ell_2 j}$ (lower row) for $1 \le \ell_3 \le 5$ computed by Eqs. (7.15) and (7.16). P-values computed by categorical regression. Upper, 1st: 4.07×10^{-11}, 2nd: 4.36×10^{-22}, 3rd: 2.03×10^{-23}, 4th: 4.14×10^{-4}, 5th: 1.57×10^{-4}. Lower, 1st: 3.36×10^{-27}, 2nd: 3.91×10^{-13}, 3rd: 7.39×10^{-9}, 4th: 9.32×10^{-5}, 5th: 2.82×10^{-5}

Finally, in order to emphasize the superiority of TD based unsupervised FE to conventional supervised methods, we apply categorical regression analysis to mRNAs expression,

$$x_{i_1 j} = a_{i_1} + \sum_s b_{i_1 s} \delta_{j s} \tag{7.19}$$

where a_{i_1} and $b_{i_1 s}$ are the regression coefficients. Based upon the results by categorical regression analysis, because too many 16,917 mRNAs probes are associated with adjusted P-values less than 0.01, we instead upload top ranked 500 mRNAs with smaller P-values to DAVID. As a result, only one GO CC enrichment, cytoplasm, associated with adjusted P-values less than 0.05, 1.9×10^{-3}, is detected. Although more advanced methods than categorical regression might achieve better performance, this drastic decrease of the number of detected GO terms enrichment demonstrates the superiority over conventional supervised method. In this sense, TD based unsupervised FE is outstanding, no matter which of type I or type II tensor is used.

Table 7.20 GO BP, CC and MF enrichment by the 374 ensembl genes identified by TD based unsupervised FE for type II tensor

Category	Term	Genes count	%	P-value	Adjusted P-value
GOTERM_BP_DIRECT	Response to estradiol	15	5.1	2.50×10^{-10}	5.00×10^{-7}
GOTERM_BP_DIRECT	Collagen catabolic process	11	3.7	8.20×10^{-8}	5.50×10^{-5}
GOTERM_BP_DIRECT	Skeletal system development	15	5.1	5.60×10^{-8}	5.70×10^{-5}
GOTERM_BP_DIRECT	Positive regulation of cell proliferation	26	8.8	2.10×10^{-7}	1.10×10^{-4}
GOTERM_BP_DIRECT	Collagen fibril organization	8	2.7	2.90×10^{-6}	1.20×10^{-3}
GOTERM_BP_DIRECT	Extracellular matrix organization	15	5.1	4.40×10^{-6}	1.30×10^{-3}
GOTERM_BP_DIRECT	Extracellular matrix disassembly	10	3.4	4.10×10^{-6}	1.40×10^{-3}
GOTERM_BP_DIRECT	Ossification	10	3.4	6.20×10^{-6}	1.60×10^{-3}
GOTERM_BP_DIRECT	Signal transduction	40	13.5	1.50×10^{-5}	3.30×10^{-3}
GOTERM_BP_DIRECT	Cell–cell signaling	15	5.1	8.00×10^{-5}	1.50×10^{-2}
GOTERM_BP_DIRECT	Response to peptide hormone	7	2.4	7.60×10^{-5}	1.50×10^{-2}
GOTERM_BP_DIRECT	Regulation of branching involved in prostate gland morphogenesis	4	1.40	1.40×10^{-4}	2.20×10^{-2}
GOTERM_BP_DIRECT	Mammary gland alveolus development	5	1.7	1.40×10^{-4}	2.40×10^{-2}
GOTERM_BP_DIRECT	Cellular response to hypoxia	9	3.0	1.80×10^{-4}	2.50×10^{-2}
GOTERM_BP_DIRECT	Immune response	19	6.4	2.10×10^{-4}	2.80×10^{-2}
GOTERM_BP_DIRECT	Proteolysis	21	7.1	2.30×10^{-4}	2.80×10^{-2}
GOTERM_BP_DIRECT	Aging	11	3.7	4.00×10^{-4}	4.60×10^{-2}
GOTERM_CC_DIRECT	Extracellular space	89	30.1	6.10×10^{-33}	1.80×10^{-30}
GOTERM_CC_DIRECT	Extracellular region	80	27.0	2.60×10^{-21}	3.90×10^{-19}
GOTERM_CC_DIRECT	Extracellular matrix	27	9.1	8.60×10^{-13}	8.60×10^{-11}
GOTERM_CC_DIRECT	Extracellular exosome	91	30.7	1.90×10^{-12}	1.40×10^{-10}
GOTERM_CC_DIRECT	Proteinaceous extracellular matrix	21	7.1	7.00×10^{-9}	4.10×10^{-7}

(continued)

Table 7.20 (continued)

Category	Term	Genes count	%	P-value	Adjusted P-value
GOTERM_CC_DIRECT	Cell surface	24	8.1	1.20×10^{-5}	6.20×10^{-4}
GOTERM_CC_DIRECT	Basement membrane	8	2.7	2.20×10^{-4}	9.20×10^{-3}
GOTERM_CC_DIRECT	Cytosol	75	25.3	4.20×10^{-4}	1.50×10^{-2}
GOTERM_MF_DIRECT	Growth factor activity	12	4.1	7.60×10^{-5}	6.70×10^{-3}
Category	Term	Genes count	%	P-value	Adjusted P-value
GOTERM_MF_DIRECT	Heparin binding	12	4.1	6.80×10^{-5}	7.20×10^{-3}
GOTERM_MF_DIRECT	Collagen binding	8	2.7	5.50×10^{-5}	7.20×10^{-3}
GOTERM_MF_DIRECT	Calcium ion binding	28	9.5	5.30×10^{-5}	9.30×10^{-3}
GOTERM_MF_DIRECT	Protein binding	178	60.1	3.90×10^{-5}	1.00×10^{-2}
GOTERM_MF_DIRECT	RAGE receptor binding	5	1.7	2.10×10^{-5}	1.10×10^{-2}
GOTERM_MF_DIRECT	Protein homodimerization activity	26	8.8	4.40×10^{-4}	3.20×10^{-2}
GOTERM_MF_DIRECT	Identical protein binding	26	8.8	6.30×10^{-4}	3.20×10^{-2}
GOTERM_MF_DIRECT	Serine-type peptidase activity	7	2.4	5.70×10^{-4}	3.30×10^{-2}
GOTERM_MF_DIRECT	Insulin-like growth factor I binding	4	1.4	8.60×10^{-4}	3.40×10^{-2}
GOTERM_MF_DIRECT	Extracellular matrix structural constituent	7	2.4	8.00×10^{-4}	3.40×10^{-2}
GOTERM_MF_DIRECT	Metalloendopeptidase activity	9	3.0	5.50×10^{-4}	3.60×10^{-2}
GOTERM_MF_DIRECT	Fibronectin binding	5	1.7	8.00×10^{-4}	3.80×10^{-2}
GOTERM_MF_DIRECT	Serine-type endopeptidase activity	13	4.4	1.10×10^{-3}	3.90×10^{-2}
GOTERM_MF_DIRECT	Protein kinase binding	16	5.4	1.40×10^{-3}	4.90×10^{-2}

Adjusted P-values are by BH criterion

7.6.2 Temporally Differentially Expressed Genes

Although type I and type II tensor approaches achieved good performance in integrated analysis of multi-class multi-omics data set in the previous section, it is better if we can demonstrate yet another example to which TD based unsupervised FE can achieve better performance. In this subsection, we try to identify genes temporally expressed distinctly between two classes.

The first data set analyzed is the comparison of NSCLC cell line H1975, with and without EGF treatment [2]. EGF is a gene supposed to accelerate cell growth and is known to be expressive frequently in cancers. Thus, EGF treatment is expected to activate cancer cell lines. The data set is composed of two mRNA expression profile, $x^{\text{control}}_{ij_1} \in \mathbb{R}^{39937\times 13}$ and $x^{\text{EGF}}_{ij_2} \in \mathbb{R}^{39937\times 15}$, which are gene expressions of cell lines without and with EGF treatment, respectively. j_1 and j_2 represent time points after the treatment (Table 7.21). Because they share genes, $x^{\text{control}}_{ij_1}$ and $x^{\text{EGF}}_{ij_2}$ can be converted to case II type I tensor as

$$x_{ij_1j_2} = x^{\text{control}}_{ij_1} x^{\text{EGF}}_{ij_2}. \tag{7.20}$$

HOSVD, Fig. 3.8, is applied to $x_{ij_1j_2}$ as

$$x_{ij_1j_2} = \sum_{\ell_1=1}^{13}\sum_{\ell_2=1}^{15}\sum_{\ell_3=1}^{39937} G(\ell_1,\ell_2,\ell_3) u^{(j_1)}_{\ell_1 j_1} u^{(j_2)}_{\ell_2 j_2} u^{(i)}_{\ell_3 i} \tag{7.21}$$

At first, we need to find singular value vectors $\boldsymbol{u}^{(j_1)}_{\ell_1} \in \mathbb{R}^{13}$ and $\boldsymbol{u}^{(j_2)}_{\ell_2} \in \mathbb{R}^{15}$ that exhibit distinct temporal expression between them. Figure 7.10 shows time development of $\boldsymbol{u}^{(j_1)}_{\ell_1}$ and $\boldsymbol{u}^{(j_2)}_{\ell_2}$ for $\ell_1 = \ell_2 = 1, 2$. Here the components of singular value vectors sharing the time points are averaged within individual vectors, $u^{(j_1)}_{\ell_1 j_1}$. It is obvious that $\boldsymbol{u}^{(j_1)}_1$ and $\boldsymbol{u}^{(j_2)}_1$ do not exhibit any time dependence while $\boldsymbol{u}^{(j_1)}_2$ and $\boldsymbol{u}^{(j_2)}_2$ do. Thus, there is a possibility that genes associated with $\boldsymbol{u}^{(j_1)}_2$ and $\boldsymbol{u}^{(j_2)}_2$ also exhibit the temporal difference between control and EGF treated cells.

In order to select genes associated with $\boldsymbol{u}^{(j_1)}_2$ and $\boldsymbol{u}^{(j_2)}_2$, we need to find $G(\ell_1,\ell_2,\ell_3)$, $\ell_1 = 2$ or $\ell_2 = 2$ having larger absolute values; $G(2,1,2)$ and $G(1,2,2)$ have larger absolute values (Table 7.22). Thus we decide to use $\boldsymbol{u}^{(i)}_2$ for FE. P-values are attributed to i as

Table 7.21 List of samples in EGF treatment experiments

Time points (h)	0	0.5	1	2	4	6	8	12	18	24	48
Control	3	1	1	1	1	1	0	0	0	2	3
EGF treated	0	2	1	1	1	1	1	1	1	3	3

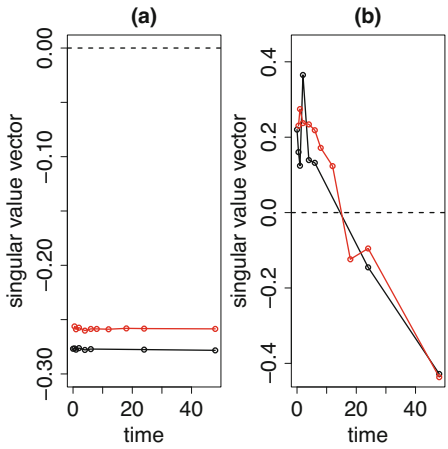

Fig. 7.10 Singular value vectors, Eq. (7.21). (**a**) $\boldsymbol{u}_1^{(j_1)}$ (black) and $\boldsymbol{u}_1^{(j_2)}$ (red). (**b**) $\boldsymbol{u}_2^{(j_1)}$ (black) and $\boldsymbol{u}_2^{(j_2)}$(red)

Table 7.22 Top ranked 10 $G(\ell_1, \ell_2, \ell_3)$s with larger absolute values among in Eq. (7.21)

ℓ_1	1	2	1	3	1
ℓ_2	1	1	2	1	3
ℓ_3	1	2	2	3	4
$G(\ell_1, \ell_2, \ell_3)$	-4.03×10^4	-1.56×10^3	1.49×10^3	1.05×10^3	-5.79×10^2
ℓ_1	4	2	5	1	4
ℓ_2	1	1	1	4	1
ℓ_3	5	3	6	6	4
$G(\ell_1, \ell_2, \ell_3)$	4.24×10^2	4.16×10^2	3.25×10^2	3.19×10^2	-2.62×10^2

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{2i}^{(i)}}{\sigma_2} \right)^2 \right]. \tag{7.22}$$

P-values are corrected by BH criterion and genes associated with adjusted P-values less than 0.01 are selected. Then 552 mRNA probes are selected.

Next, we need to see if the selected 552 mRNA probes really exhibit temporal difference between control and EGF treated cells. For this purpose, we compute correlation coefficient between

$$\left(x_{i1}^{\text{control}}, \ldots, x_{i13}^{\text{control}}, x_{i1}^{\text{EGF}}, \ldots, x_{i15}^{\text{EGF}} \right) \tag{7.23}$$

and

$$\left(u_{2,1}^{(j_1)}, \ldots, u_{2,13}^{(j_1)}, u_{2,1}^{(j_2)}, \ldots, u_{2,15}^{(j_2)} \right) \tag{7.24}$$

to see if 552 selected genes are coincident with $\boldsymbol{u}_2^{(j_1)}$ and $\boldsymbol{u}_2^{(j_2)}$. Figure 7.11a shows the histogram of correlation coefficients. Because there are two peaks at ± 1, it is

obvious that gene expression of selected 552 mRNA probes is highly coincident with $\boldsymbol{u}_2^{(j_1)}$ and $\boldsymbol{u}_2^{(j_2)}$.

Before comparing 552 genes directly between control and EGF treated cells, we need shift and scale individual gene expression profiles such that they have same baseline and amplitude. In order that, we apply the following linear regression

$$u_{2j_1}^{(j_1)} = a_i x_{ij_1}^{\text{control}} + b_i \tag{7.25}$$

$$u_{2j_2}^{(j_2)} = a_i x_{ij_2}^{\text{EGF}} + b_i \tag{7.26}$$

where a_i and b_i are the regression coefficients. Because regression coefficients are shared between control and EGF treated ones, this does not reduce the difference between these two. Then, we compare $a_i x_{ij_1}^{\text{control}} + b_i$ and $a_i x_{ij_2}^{\text{EGF}} + b_i$ of selected 552 mRNA probes (Fig. 7.11b). Not all, but the comparisons of five out of seven time points excluding two time points, 4 and 24 h, after the EGF treatment are associated with P-values less than 0.05. Thus, TD based unsupervised FE has the ability to select genes associated with temporal distinction.

Next, we try to see if type II tensor approach works as well. Because case II tensor share the feature whose number is generally much larger than the number of samples, type II tensor where shared dimension is summed up can result in much smaller number of components. Type II tensor is defined as

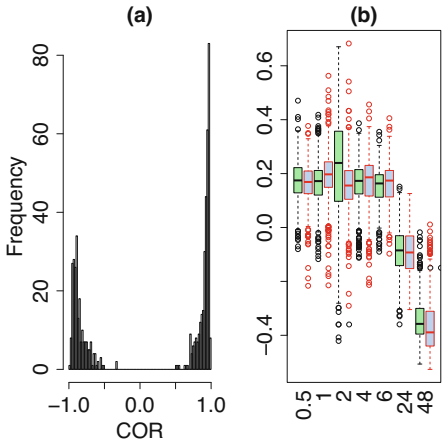

Fig. 7.11 (**a**) Histogram of correlation coefficients between Eqs. (7.23) and (7.24) for case II type I tensor, Eq. (7.20). (**b**) Boxplot of Eqs. (7.25) (black boxes filled with green) and (7.26) (red boxes filled with blue) for case II type I tensor, Eq. (7.20). P-values computed by t test: 0.5 h:2.83×10^{-2}, 1 h:6.81×10^{-8}, 2 h:5.63×10^{-12}, 4 h:3.5×10^{-1}, 6 h:4.83×10^{-2}, 24 h:5.0×10^{-1}, 48 h:1.70×10^{-6}

$$x_{j_1 j_2} = \sum_{i=1}^{39937} x_{i j_1 j_2}. \tag{7.27}$$

where $x_{i j_1 j_2}$ is defined in Eq. (7.20). The number of components in $x_{j_1 j_2} \in \mathbb{R}^{13\times 15}$ is $13 \times 15 = 195$, which is as small as 1/39937 of the number of components in $x_{i j_1 j_2} \in \mathbb{R}^{39937\times 13\times 15}$. Thus, if type II tensor approach works as well, it is very effective. SVD is applied to $x_{j_1 j_2}$ as

$$x_{j_1 j_2} = \sum_{\ell} \lambda_\ell u_{\ell j_1}^{(j_1)} u_{\ell j_2}^{(j_2)} \tag{7.28}$$

Figure 7.12 shows the $\boldsymbol{u}_\ell^{(j_1)}$ and $\boldsymbol{u}_\ell^{(j_2)}$ for $\ell = 1, 2$. Basically, it looks similar to Fig. 7.10. Thus we decide to employ $\ell = 2$ for FE. Then, singular value vectors attributed to i can be computed as Eq. (5.14),

$$u_{\ell i}^{i;j_1} = \sum_{j_1=1}^{13} x_{i j_1}^{\text{control}} u_{\ell j_1}^{(j_1)} \tag{7.29}$$

$$u_{\ell i}^{i;j_2} = \sum_{j_2=1}^{15} x_{i j_2}^{\text{EGF}} u_{\ell j_2}^{(j_2)} \tag{7.30}$$

Thus P-values are also attributed to i in two ways as

$$P_i^{j_1} = P_{\chi^2}\left[> \left(\frac{u_{2i}^{(i;j_1)}}{\sigma_2} \right)^2 \right], \tag{7.31}$$

$$P_i^{j_2} = P_{\chi^2}\left[> \left(\frac{u_{2i}^{(i;j_2)}}{\sigma_2'} \right)^2 \right]. \tag{7.32}$$

P-values are corrected by BH criterion. mRNA probes associated with adjusted P-values less than 0.01 are selected. Then, 482 and 487 mRNA probes, between which 396 mRNA probes are chosen in common, are selected using $P_i^{j_1}$ and $P_i^{j_2}$, respectively. Thus, in some sense, type II tensor approach can give the results coincident between two approximations of singular value vectors attributed to i using Eqs. (7.29) and (7.30), respectively.

Next, we need to see if the 396 mRNA probes chosen in common really exhibit temporal difference between control and EGF treated cells as in the case of type I tensor approach. The correlation coefficient between Eqs. (7.23) and (7.24) is computed again to see the coincidence between gene expression and singular value vectors (Fig. 7.13a). It is obvious that the peaks at ± 1 is much steeper than that in

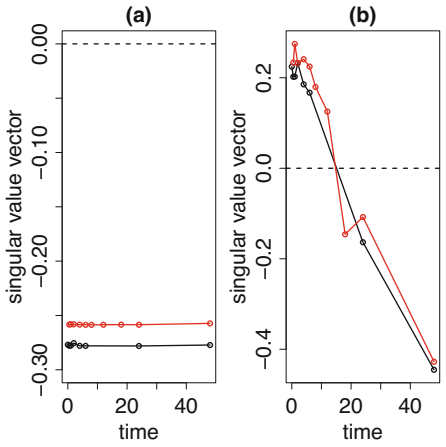

Fig. 7.12 Singular value vectors, Eq. (7.28). (**a**) $\boldsymbol{u}_1^{(j_1)}$ (black) and $\boldsymbol{u}_1^{(j_2)}$ (red). (**b**) $\boldsymbol{u}_2^{(j_1)}$ (black) and $\boldsymbol{u}_2^{(j_2)}$(red)

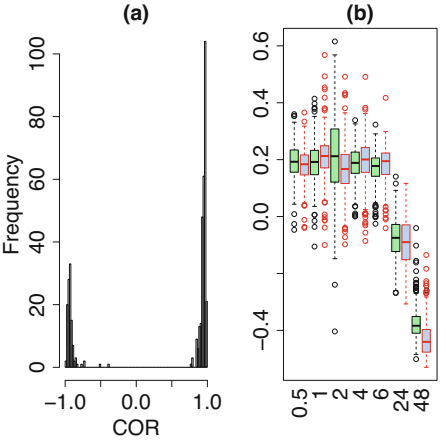

Fig. 7.13 (**a**) Histogram of correlation coefficients between Eqs. (7.23) and (7.24) for case II type II tensor, Eq. (7.27). (**b**) Boxplot of Eqs. (7.25) (black boxes filled with green) and (7.26) (red boxes filled with blue) for case II type II tensor, Eq. (7.27). P-values computed by t test: 0.5 h:1.68×10^{-2}, 1 h:2.56×10^{-5}, 2 h: 3.83×10^{-7}, 4 h:9.14×10^{-2}, 6 h:7.30×10^{-4}, 24 h:2.36×10^{-2}, 48 h:5.55×10^{-38}

Fig. 7.11a. This suggests that type II tensor approach might be better than type I tensor approach in spite of the smaller computational resources required.

In order to confirm the superiority of type II tensor approach, we again apply linear regression Eqs. (7.25) and (7.26) replacing singular value vectors with those obtained by type II tensor (Fig. 7.13b). Because six among seven time points excluding 4 h after the EGF treatment are associated with P-values less than 0.05, type II tensor approach is superior to type I tensor approach.

Finally, in order to validate 552 and 396 mRNA probes selected by type I and II tensor approaches, respectively, we upload RefSeq mRNA IDs associated with these probes to DAVID. Table 7.23 lists the KEGG pathways identified by DAVID for type I and II tensor approach. Although common five KEGG pathways are associated with adjusted P-values less than 0.05, P-values for type II tensor approach are smaller than those for type I tensor approach. Because P-values are more likely smaller for more number of genes uploads, smaller P-values attributed to KEGG

Table 7.23 KEGG pathways identified by DAVID for genes associated with 552 (upper numbers) and 396 (lower numbers) miRNA probes selected using type I, Eq. (7.20), and II, Eq. (7.27), tensor approach

Category	Term	Count	%	P-value	Adjusted P-value
KEGG_PATHWAY	Cell cycle	29	9.0	7.2×10^{-24}	1.0×10^{-21}
		28	12.1	3.7×10^{-29}	3.2×10^{-27}
KEGG_PATHWAY	Oocyte meiosis	14	4.3	7.6×10^{-8}	5.5×10^{-6}
		14	6.0	1.4×10^{-10}	5.8×10^{-9}
KEGG_PATHWAY	DNA replication	8	2.5	2.8×10^{-6}	1.4×10^{-4}
		9	3.9	3.2×10^{-9}	9.3×10^{-8}
KEGG_PATHWAY	Progesterone-mediated oocyte maturation	8	2.5	9.2×10^{-4}	3.3×10^{-2}
		9	3.9	4.0×10^{-6}	8.6×10^{-5}
KEGG_PATHWAY	p53 signaling pathway	7	2.2	1.2×10^{-3}	3.5×10^{-2}
		6	2.6	7.7×10^{-4}	1.3×10^{-2}

Adjusted P-values are by BH criterion

pathways by type II tensor approach where less number of genes are selected suggest the superiority of type II tensor approach from the biological point of view.

Although type II approach is better than type I approach in this specific example, because it is highly dependent upon data sets analyzed, it is difficult to know in advance which is better.

7.7 Gene Expression and Methylation in Social Insects

As the first example of the application of case I tensor approach, we employ the multi-omics analysis of social insects. Social insects, e.g., ants and bees, are known to have castes where distinct phenotypes appear in spite of shared genome. Thus, it is interesting to know what drives differentiation between castes.

One possible scenario is the alteration of epigenome [29], because epigenome has plasticity that can mediate differentiation between castes. Most typical caste is composed of queen and worker. The former, queen, concentrates on reproduction while the latter, workers, serve to maintain colony. In spite of their strict difference of phenotype, they are often known to be relatives. Thus, they share genome to some extent with having distinct phenotype. This suggests that epigenome can play potential roles in the differentiation of caste.

In this section, we try to identify genes associated with differential expression and methylation between caste, especially queens and workers [25], because such genes are potential candidates that can mediate distinct phenotypes between castes. In order that, we employ TD based unsupervised FE that can integrate multi-omics data sets. The data set analyzed [16] is composed of two insect species, bee (*P.*

Table 7.24 Number of samples in social insect study [16]

	Methylation			mRNA	
Caste	Control	Queen	Worker	Queen	Worker
P. canadensis	1	3	3	4	6
D. quadriceps	1	3	3	7	6

canadensis) and ant (*D. quadriceps*). Table 7.24 shows the number of samples available from GEO with GEO ID GSE59525. As can be seen, it is a typical large p small n data set.

Because the amount of gene expression is measured by the unit of Reads Per Kilobase of exon per Million mapped reads (RPKM), it is used as it is. Because the gene expression profile of *P. canadensis* was $\log_2$-ratio converted, it is expanded to the original one as 2^x where x is gene expression. On the other hand, we would like to employ case II tensor format (Table 5.3) where genes are shared. Thus we need to convert methylation profiles to be attributed to individual genes. In order that, assuming m_{s_1} and m_{s_2} are methylation and nonmethylation values, respectively, at locus s, then the relative methylation within the ith gene can be defined as

$$\frac{\sum_{s\in i} m_{s_1}}{\sum_{s\in i}\left(m_{s_1}+m_{s_2}\right)} \tag{7.33}$$

where $\sum_{s\in i}$ is taken over s bases within DNA sequences corresponding to the ith gene body; the reason why methylation not in promoter region but in the gene body is summed up and is attributed to genes is because gene body methylation is believed to affect gene expression in insects [32]. Relative methylation profile is formatted as

$$x_{ik}^{\text{metyl, bee}} \in \mathbb{R}^{N\times 7}, \tag{7.34}$$

$$x_{ik}^{\text{metyl, ant}} \in \mathbb{R}^{N\times 7}, \tag{7.35}$$

where N is the number of genes. $k=1$ corresponds to control samples. $2\le k\le 4$ and $5\le k\le 7$ correspond to queens and workers, respectively. On the other hand, mRNA expression is formatted as

$$x_{ij}^{\text{mRNA, bee}} \in \mathbb{R}^{N\times 10}, \tag{7.36}$$

$$x_{ij}^{\text{mRNA, ant}} \in \mathbb{R}^{N\times 13}. \tag{7.37}$$

where $1\le j\le 4$ and $5\le j\le 10$ for bee correspond to queens and workers, respectively, while $1\le j\le 7$ and $8\le j\le 13$ for ant correspond to queens and workers, respectively. Then case II tensor is generated as

$$x_{ijk}^{\text{bee}} = x_{ij}^{\text{mRNA, bee}} x_{ik}^{\text{metyl, bee}}, \tag{7.38}$$

$$x^{\text{ant}}_{ijk} = x^{\text{mRNA, ant}}_{ij} x^{\text{metyl, ant}}_{ik}, \tag{7.39}$$

where $x^{\text{bee}}_{ijk} \in \mathbb{R}^{N\times 10\times 7}$ and $x^{\text{ant}}_{ijk} \in \mathbb{R}^{N\times 13\times 7}$. HOSVD, Fig. 3.8, is applied to x^{bee}_{ijk} and x^{ant}_{ijk} as

$$x^{\text{bee}}_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{10}\sum_{\ell_3=1}^{7} G(\ell_1,\ell_2,\ell_3) u^{\text{bee}(i)}_{\ell_1 i} u^{\text{bee}(j)}_{\ell_2 j} u^{\text{bee}(k)}_{\ell_3 k} \tag{7.40}$$

$$x^{\text{ant}}_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{13}\sum_{\ell_3=1}^{7} G(\ell_1,\ell_2,\ell_3) u^{\text{ant}(i)}_{\ell_1 i} u^{\text{ant}(j)}_{\ell_2 j} u^{\text{ant}(k)}_{\ell_3 k} \tag{7.41}$$

where $u^{\text{bee}(i)}_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $u^{\text{bee}(j)}_{\ell_2 j} \in \mathbb{R}^{10\times 10}$, $u^{\text{bee}(k)}_{\ell_3 k} \in \mathbb{R}^{7\times 7}$, $u^{\text{ant}(i)}_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $u^{\text{ant}(j)}_{\ell_2 j} \in \mathbb{R}^{13\times 13}$, and $u^{\text{ant}(k)}_{\ell_3 k} \in \mathbb{R}^{7\times 7}$.

Next, as usual, we need to find which singular value vectors are coincident with the distinction between queens and workers. Figures 7.14a and b, 7.15a and b show singular value vectors associated with highest distinction between queens and workers. Unfortunately, singular value vectors of methylation do not exhibit small enough P-values to be significant. Nevertheless, because selected genes might exhibit significant distinct expression between queens and workers, we continue the procedure. We seek $G(\ell_1, 1, 3)$ for *P. canadensis* and $G(\ell_1, 1, 5)$ for *D. quadriceps* with larger absolute values.

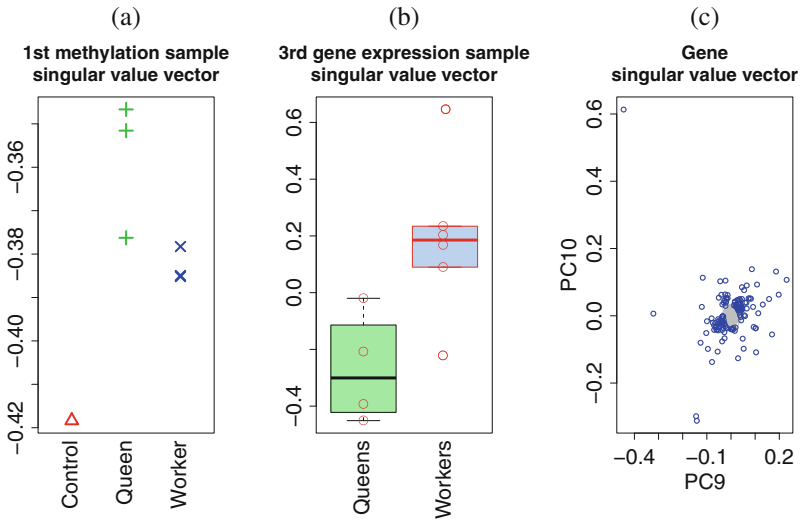

Fig. 7.14 Singular value vectors for *P. canadensis*. P-values are computed by t test between queens and workers. (**a**) $\boldsymbol{u}^{\text{bee}(k)}_1$, $P = 1.1\times 10^{-1}$ (**b**) $\boldsymbol{u}^{\text{bee}(j)}_3$, $P = 1.65\times 10^{-2}$ (**c**) $\boldsymbol{u}^{\text{bee}(i)}_{\ell_1}$, $\ell_1 = 9, 10$. Blue open circles are selected genes

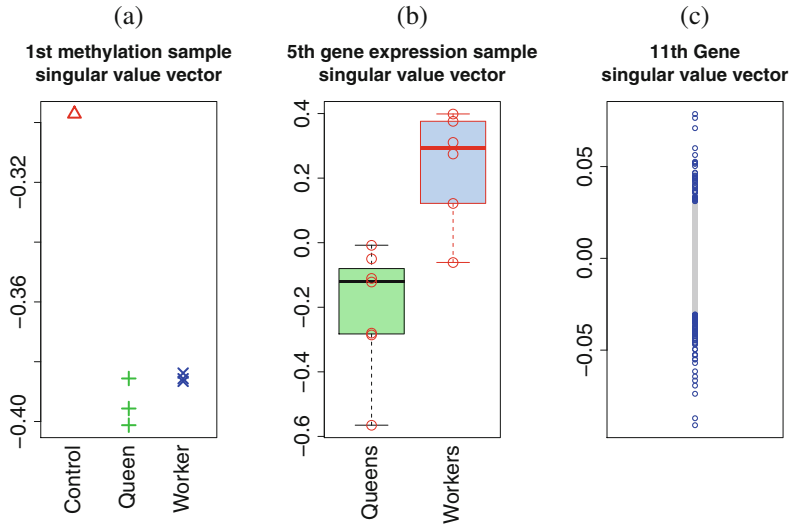

Fig. 7.15 Singular value vectors for *D. quadriceps*. P-values are computed by t test between queens and workers. (**a**) $\boldsymbol{u}_1^{\text{ant}(k)}$, $P = 1.9 \times 10^{-1}$ (**b**) $\boldsymbol{u}_5^{\text{ant}(j)}$, $P = 1.25 \times 10^{-3}$ (**c**) $\boldsymbol{u}_{11}^{\text{ant}(i)}$. Blue open circles are selected genes

Table 7.25 The top 10 core tensors, G, with large absolute values

P. canadensis		*D. quadriceps*	
ℓ_1	$G(\ell_1, 1, 3)$	ℓ_1	$G(\ell_1, 1, 5)$
9	−79.8	11	−54.8
10	75.4	12	4.1
7	−61.4	25	3.4
11	38.4	2	−2.9
5	−23.4	23	2.8
4	−16.0	9	2.4
12	−11.9	20	−2.2
1	−5.4	8	2.2
13	5.4	10	−1.7
6	−4.5	22	−1.4

Table 7.25 lists the top ranked Gs with larger absolute values. Then we decide that $\boldsymbol{u}_{\ell_1}^{\text{bee}(i)}$, $\ell_1 = 9, 10$ and $\boldsymbol{u}_{11}^{\text{ant}(i)}$ are used for FE (Figs. 7.14c and 7.15c). P-values are attributed to ith gene as

$$P_i^{\text{bee}} = P_{\chi^2}\left[> \sum_{\ell_1=9}^{10} \left(\frac{u_{\ell_1 i}^{\text{bee}(i)}}{\sigma_{\ell_1}} \right)^2 \right], \tag{7.42}$$

and

Table 7.26 Statistical tests of the differences (between queens and workers) in gene expression and methylation

		t	Wilcox	KS
P. canadensis	Gene expression	1.71×10^{-3}	1.89×10^{-2}	0.08
	Methylation	1.74×10^{-4}	5.06×10^{-3}	1.02×10^{-3}
D. quadriceps	Gene expression	2.73×10^{-12}	9.05×10^{-12}	4.41×10^{-11}
	Methylation	0.3757	0.7163	0.4413

The genes identified by TD-based unsupervised FE are analyzed by t (the t test), Wilcox (the Wilcoxon rank sum test), and KS (the Kolmogorov–Sinai test), all two-sided

$$P_i^{\text{ant}} = P_{\chi^2}\left[> \left(\frac{u_{11i}^{\text{ant}(i)}}{\sigma_{11}} \right)^2 \right], \tag{7.43}$$

P-values are adjusted by BH criterion. Genes associated with adjusted P-values less than 0.01 are selected. As a result, 133 and 128 genes are selected for *P. canadensis* and *D. quadriceps*, respectively.

The point is if selected genes are associated with distinct gene expression and methylation between queens and workers simultaneously. Then we apply three statistical tests to 133 genes and 128 genes between queens and workers (Table 7.26). Selected genes exhibit simultaneous distinct gene expression and methylation between queens and workers for *P. canadensis*, but not for *D. quadriceps*. Thus selected genes can be potential factors that can mediate caste differentiation for *P. canadensis*, but not for *D. quadriceps*. Although we are not sure the lack of detection for *D. quadriceps* is because of biological reason or failure of our methodology, at least, our purpose is achieved for *P. canadensis*. In order to clarify this point, we need to continue research.

In order to see if conventional supervised methods can do this, we apply t test to gene expression and promoter methylation to find genes that exhibit significant distinction between queens and workers. As a result, two genes for distinct gene expression between queens and workers for *D. quadriceps* are associated with adjusted P-vales less than 0.01. This poor performance is because of small number of samples. Thus, TD based unsupervised FE has the ability to find significant genes for large p small n problem, for which conventional supervised method fails.

Before closing this section, we would like to validate selected genes from the biological point of view. Because these two insects are not included in popular enrichment servers, e.g. DAVID or Enrichr, instead we download list of GO terms,[1] PCAN.v01.GO.tsv for *P. canadensis* and DQUA.v01.GO.tsv for *D. quadriceps*. Fisher's exact test is performed in order to evaluate enrichment and computed P-values are corrected by BH criterion. GO terms associated with adjusted P-values less than 0.05 are searched. There are three GO terms, Lipid transporter activity

[1] Paper Wasp and Dinosaur Ant Project. Accessed 15 Jan. 2019. http://wasp.crg.eu/download.html.

(GO:0005319), Lipid particle (GO:0005811), and Lipid transport (GO:0006869) enriched in 133 genes selected for *P. canadensis*, while there are no GO terms enriched in 128 genes selected for *D. quadriceps*. This might be reasonable because 128 genes selected for *D. quadriceps* are not associated with distinct methylation between queens and workers (Table 7.26). Anyway, 133 genes selected for *P. canadensis*, which is simultaneously associated with distinct gene expression and methylation between queens and workers, are associated with a few GO term enrichment. Thus, at least for *P. canadensis*, TD based unsupervised FE is useful also from the biological point of view.

7.8 Drug Discovery From Gene Expression: II

In Sect. 7.3, we have already shown that TD based unsupervised FE successfully identifies compounds that affect gene expression in dose-dependent manner and their target proteins from only gene expression profiles in fully unsupervised manner. Nevertheless, it is strictly restricted to cancers because gene expression profiles are measured in cancer cell lines. The identifying drug compounds that are effective to other diseases requires additional gene expression profiles treated by compounds in specific diseases, e.g., model animals or cell lines originated from the disease. Thus in the manner in Sect. 7.3, the effectiveness of methods is quite limited.

In this section, with using case II tensor where genes are shared between two matrices or tensors, we try to identify disease effective drugs without measuring gene expression repeatedly for individual diseases. The study design is as follows (Fig. 7.16). $x_{ij_1j_2}$ is the ith gene expression profiles of animals treated by j_1 compound at the time point j_2 after the treatment. x_{ij_3} is the human gene expression profile of gene i at j_3th patients or healthy control. Case II tensor $x_{ij_1j_2j_3}$ is generated as

$$x_{ij_1j_2j_3} = x_{ij_1j_2}x_{ij_3} \tag{7.44}$$

HOSVD algorithm, Fig. 3.8, is applied to $x_{ij_1j_2j_3}$ as

$$x_{ij_1j_2j_3} = \sum_{\ell_1=1}^{N_1}\sum_{\ell_2=1}^{N_2}\sum_{\ell_3=1}^{N_3}\sum_{\ell_4=1}^{N_4} G(\ell_1,\ell_2,\ell_3,\ell_4)u^{(j_1)}_{\ell_1 j_1}u^{(j_2)}_{\ell_2 j_2}u^{(j_3)}_{\ell_3 j_3}u^{(i)}_{\ell_4 i} \tag{7.45}$$

Then, $\boldsymbol{u}^{(j_2)}_{\ell_2}$ that exhibits time dependence and $\boldsymbol{u}^{(j_3)}_{\ell_3}$ that exhibits distinction between healthy controls and patients are searched. After identifying ℓ_2 and ℓ_3, ℓ_1 and ℓ_4 associated with $G(\ell_1,\ell_2,\ell_3,\ell_4)$ with larger absolute values are selected. Once, ℓ_1 and ℓ_4 are selected, P-values are attributed to i and j_1 as

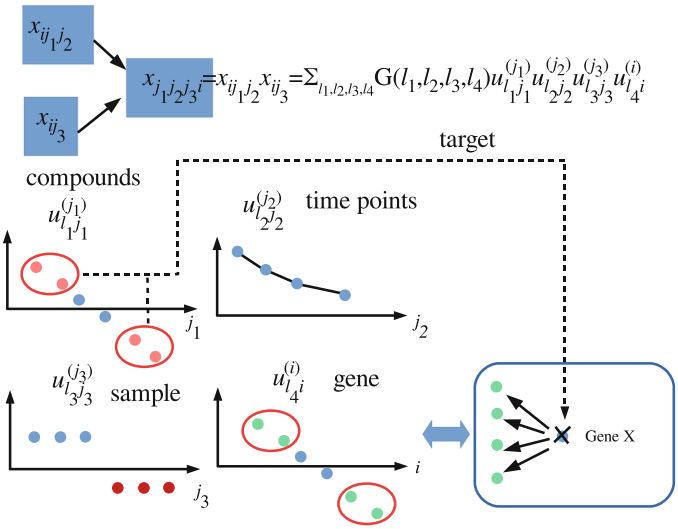

Fig. 7.16 Integrated analysis of gene expression profile of drug treated animals, $x_{ij_1j_2}$ and human gene expression profiles of patients and healthy control, x_{ij_3}. i: genes, j_1: compounds, j_2: time point after the treatment, j_3: human samples

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}} \right)^2 \right], \tag{7.46}$$

and

$$P_{j_1} = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 j_1}}{\sigma_{\ell_1}} \right)^2 \right]. \tag{7.47}$$

P-values are corrected by BH criterion and i and j_1 associated with adjusted P-values less than 0.01 (filled pink circles and filled light green circles surrounded by pink oval in Fig. 7.16) are supposed to be selected. Target proteins are decided by the comparison with external databases (as shown in Fig. 7.5). This process results in the set of drug candidates compounds and candidate target proteins. Figure 7.17 and Table 7.27 summarize the process till selection of singular value vectors attributed to genes and compounds. There are six diseases analyzed: heart failure, PTSD, acute lymphoblastic leukemia (ALL), diabetes, renal carcinoma, and cirrhosis. In some cases, modes of case II tensors are more than four because human gene expression profiles are represented as not matrices but tensors.

Gene expression profiles of model animals are downloaded from DrugMatrix [15] where rats are treated as model animals and gene expression profiles of various tissues are extracted. Corresponding human or rat disease expression profiles are downloaded from GEO. For heart failure, human disease heart failure

Table 7.27 A summary of TDs and identification of various singular value vectors for identification of candidate drugs and genes used to find genes encoding drug target proteins

Diseases	Tensors			Core tensor	Singular value vectors
	DrugMatrix	Disease	Generated		
Heart failure	$x_{ij_1j_2}$ $\in \mathbb{R}^{N_4\times N_1\times N_2}$	x_{ij_3} $\in \mathbb{R}^{N_4\times N_3}$	$x_{ij_1j_2j_3}$ $\in \mathbb{R}^{N_4\times N_1\times N_2\times N_3}$	$G(\ell_1\ell_2\ell_3\ell_4)$ $\in \mathbb{R}^{N_1\times N_2\times N_3\times N_4}$	$u^{(jk)}_{\ell_k,j_k}, k\le 3, u^{(i)}_{\ell_4,i}\in\mathbb{R}^{N_k\times N_k}$ $(N_1, N_2, N_3, N_4) = (218, 4, 313, 3937)$
Selected				$\ell_1 = 2; \ell_2 = 2; \ell_3 = 2, 3; \ell_4 = 21, 25, 27, 28, 33, 36, 37, 41, 42, 48$	
PTSD rat model	$x_{ij_1j_2}$ $\in \mathbb{R}^{N_6\times N_1\times N_2}$	$x_{ij_3j_k}, k = 4, 5$ $\in \mathbb{R}^{N_6\times N_3\times N_k}$	$x_{ij_1j_2j_3j_4j_5}$ $\in \mathbb{R}^{N_6\times N_1\times N_2\times N_3\times N_4\times N_5}$	$G(\ell_1\ell_2\ell_3\ell_4\ell_5\ell_6)$ $\in \mathbb{R}^{N_1\times N_2\times N_3\times N_4\times N_5\times N_6}$	$u^{(jk)}_{\ell_k,j_k}, k\le 5, u^{(i)}_{\ell_6,i}\in\mathbb{R}^{N_k\times N_k}$ $(N_1, N_2, N_3, N_4, N_5, N_6) = (22, 4, 2, 15, 15, 7501)$
Selected				$\ell_1 = 2; \ell_2 = 2; \ell_3 = 1; \ell_4 = \ell_5 = 3; \ell_6 = 75, 77, 81, 83, 84, 85, 89, 90, 102$	
ALL	$x_{ij_1j_2}$ $\in \mathbb{R}^{N_5\times N_1\times N_2}$	$x_{ij_3j_4}$ $\in \mathbb{R}^{N_5\times N_3\times N_4}$	$x_{ij_1j_2j_3j_4}$ $\in \mathbb{R}^{N_5\times N_1\times N_2\times N_3\times N_4}$	$G(\ell_1\ell_2\ell_3\ell_4\ell_5)$ $\in \mathbb{R}^{N_1\times N_2\times N_3\times N_4\times N_5}$	$u^{(jk)}_{\ell_k,j_k}, k\le 4, u^{(i)}_{\ell_5,i}\in\mathbb{R}^{N_k\times N_k}$ $(N_1, N_2, N_3, N_4, N_5) = (77, 4, 4, 74, 2597)$
Selected				$\ell_1 = 2, 3, 5, 6, 9, 10; \ell_2 = 3; \ell_3 = 4; \ell_5 = 1, 2, 3, 5$	
Diabetes	$x_{ij_1j_2}$ $\in \mathbb{R}^{N_4\times N_1\times N_2}$	x_{ij_3} $\in \mathbb{R}^{N_4\times N_3}$	$x_{ij_1j_2j_3}$ $\in \mathbb{R}^{N_4\times N_1\times N_2\times N_3}$	$G(\ell_1\ell_2\ell_3\ell_4)$ $\in \mathbb{R}^{N_1\times N_2\times N_3\times N_4}$	$u^{(jk)}_{\ell_k,j_k}, k\le 3, u^{(i)}_{\ell_4,i}\in\mathbb{R}^{N_k\times N_k}$ $(N_1, N_2, N_3, N_4) = (253, 4, 69, 3489)$
Selected				$\ell_1 = 2; \ell_2 = 2; \ell_3 = 1, 4; \ell_4 = 1, 4$	
Renal carcinoma	$x_{ij_1j_2}$ $\in \mathbb{R}^{N_4\times N_1\times N_2}$	x_{ij_3} $\in \mathbb{R}^{N_4\times N_3}$	$x_{ij_1j_2j_3}$ $\in \mathbb{R}^{N_4\times N_1\times N_2\times N_3}$	$G(\ell_1\ell_2\ell_3\ell_4)$ $\in \mathbb{R}^{N_1\times N_2\times N_3\times N_4}$	$u^{(jk)}_{\ell_k,j_k}, k\le 3, u^{(i)}_{\ell_4,i}\in\mathbb{R}^{N_k\times N_k}$ $(N_1, N_2, N_3, N_4) = (253, 4, 202, 4036)$
Selected				$\ell_1 = 2; \ell_2 = 2; \ell_3 = 13, 15, 30, 33, 35; \ell_4 = 186, 215, 233, 244, 251, 269, 274, 309, 312, 318$	
Cirrhosis	$x_{ij_1j_2}$ $\in \mathbb{R}^{N_4\times N_1\times N_2}$	x_{ij_3} $\in \mathbb{R}^{N_4\times N_3}$	$x_{ij_1j_2j_3}$ $\in \mathbb{R}^{N_4\times N_1\times N_2\times N_3}$	$G(\ell_1\ell_2\ell_3\ell_4)$ $\in \mathbb{R}^{N_1\times N_2\times N_3\times N_4}$	$u^{(jk)}_{\ell_k,j_k}, k\le 3, u^{(i)}_{\ell_4,i}\in\mathbb{R}^{N_k\times N_k}$ $(N_1, N_2, N_3, N_4) = (355, 4, 216, 3961)$
Selected				$\ell_1 = 2; \ell_2 = 2; \ell_3 = 2, 6; 2 \le \ell_4 \le 10$	

In all cases, ℓ_1 stands for singular value vectors of compounds, whereas ℓ_k with the last (largest) k denotes gene singular value vectors. ℓ_2 stands for singular value vectors of time points in DrugMatrix data. The remaining singular value vectors correspond to sample singular value vectors depending on the properties of gene expression profiles of diseases. See also Fig. 7.17 for the corresponding data

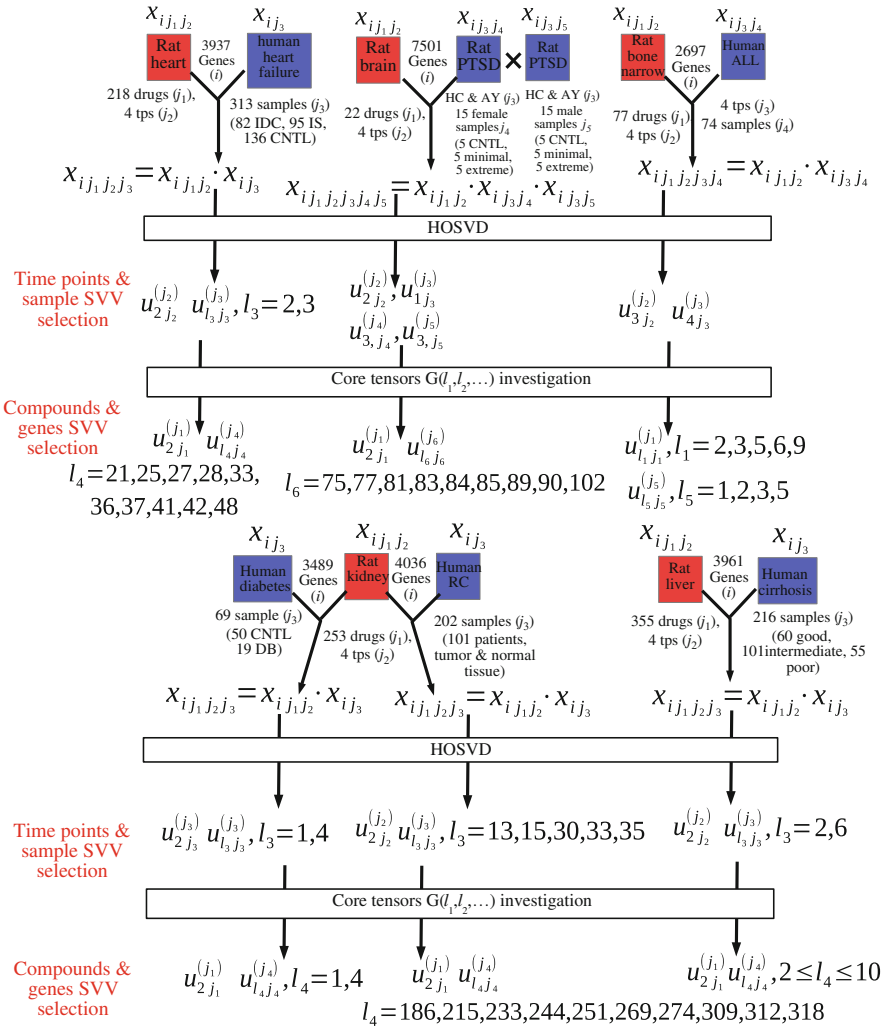

Fig. 7.17 Schematics that illustrate the procedure of TD-based unsupervised FE applied to the various disease and DrugMatrix data sets. SVV: singular value vector. Selected four time points (tps) are 1/4, 1, 3, and 5 days after treatment

gene expression profiles and rat heart gene expression profiles treated by drugs are used. For PTSD, stressed mouse brain gene expression profiles and rat brain gene expression profiles treated by drugs are used. For ALL, drug treated rat and ALL human patients bone marrow gene expression profiles are used. For diabetes and renal carcinoma, drug treated rat kidney gene expression profiles are used. Diabetes and renal carcinoma human patients kidney gene expression profiles are used for diabetes and renal carcinoma, respectively. For cirrhosis, drug treated rat

Table 7.28 The number of genes, drugs, and target proteins identified by TD based unsupervised FE

Disease	Inferred genes	Inferred compounds	Predicted target	
			Up	Down
Heart failure	274	43	556	449
PTSD	374	6	578	548
ALL	24	2	91	57
Diabetes	65	14	186	140
Renal carcinoma	225	14	229	177
Cirrhosis	132	27	510	488

liver gene expression profiles and cirrhosis human liver expression profiles are used. See appendix for more details.

After selecting genes and drugs, genes are uploaded to Enrichr for target protein identification. Genes enriched (adjusted P-values less than 0.01) in "Single gene perturbation GEO up" and "Single gene perturbation GEO down" are selected as target proteins. This process is similar to that illustrated in Fig. 7.5. Table 7.28 summarizes the number of identified genes, compounds, and target proteins.

In order to validate the relationship between drugs and target proteins predicted, we compare them with DINIES [31] that stores known protein–drug interactions. We upload drugs one by one to DENIES with parameters "chemogenomic approach" and "with learning on all DBs" and can get list of target proteins. They are merged into a list of proteins because individual proteins can be targeted by multiple drugs. The obtained set of target proteins are compared with predicted targets in Table 7.28. Here total proteins considered is limited to genes included in "Single_Gene_Perturbations_from_GEO_all_list" of Enrichr. Table 7.29 shows the results of evaluation by Fisher's exact test and χ^2 test. Ten out of twelve are evaluated as significant (P-values less than 0.05) by either Fisher's exact test or χ^2 test. This suggests that TD based unsupervised FE can be used for the prediction of target protein and diseases of drugs only from gene expression profile, in fully unsupervised manner in the sense that it does not require any pre-knowledge about disease–drug or protein–drug interaction.

7.9 Integrated Analysis of miRNA Expression and Methylation

Unsupervised method is often useful when applied to something for which no pre-knowledge is available. For example, two kinds of omics data might be correlated with unknown reasons. To search this kind of hidden (latent) relationship, unsupervised method is critically useful. In this section, we propose the application

Table 7.29 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

		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	490	1.1×10^{-2}	8.1×10^{-3}	2.91	595	467	1.6×10^{-3}	1.1×10^{-3}	3.81
	T	8	20				7	21			

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

of case I type II tensor to investigate relationship between miRNA expression and methylation, between which no direct relationships are biologically expected.

Promoter methylation of genes targeted by miRNAs can of course affect expression of these genes. Nevertheless, there seem to be no biological reasons that promoter methylation of genes targeted by miRNAs affects the expression of these miRNAs themselves or vice versa. Thus, if we can find any correlations between these two, it might be a starting point of finding new biological points of view.

In this section, we make use of TCGA data set [28]. The data set we analyze is composed of eight normal ovarian tissue samples and 569 tumor samples. Our data set includes expression data on 723 miRNAs as well as promoter methylation profiles of 24,906 genes. They are formatted as matrices

$$x_{ij}^{\text{methyl}} \in \mathbb{R}^{24906 \times 577} \tag{7.48}$$

$$x_{kj}^{\text{miRNA}} \in \mathbb{R}^{723 \times 577} \tag{7.49}$$

They are converted to case I tensor because they share samples as

$$x_{ijk} = x_{kj}^{\text{miRNA}} x_{ij}^{\text{methyl}} \tag{7.50}$$

Usually, HOSVD, Fig. 3.8, is supposed to be applied to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{24906} \sum_{\ell_2=1}^{577} \sum_{\ell_3=1}^{723} G(\ell_1, \ell_2, \ell_3) u^{(i)}_{\ell_1 i} u^{(j)}_{\ell_2 j} u^{(k)}_{\ell_3 k}. \tag{7.51}$$

Unfortunately, x_{ijk} is too huge to apply HOSVD directly. Thus, instead, we derive type II tensor as

$$x_{ik} = \sum_{j=1}^{577} x_{ijk}. \tag{7.52}$$

Now it is a matrix. Thus we can apply PCA to it. Then we can have PC score $\boldsymbol{u}_\ell \in \mathbb{R}^{723}$ attributed to miRNA and PC loading $\boldsymbol{v}_\ell \in \mathbb{R}^{24906}$ attributed to methylation. The singular value vectors attributed to sample j are computed in two ways as Eq. (5.15)

$$u^{(j;k)}_{\ell j} = \sum_k u_{\ell k} x^{\text{miRNA}}_{kj}, \tag{7.53}$$

$$u^{(j;i)}_{\ell j} = \sum_i v_{\ell i} x^{\text{methyl}}_{ij}. \tag{7.54}$$

The first thing to check is if there are any ℓs such that $\boldsymbol{u}^{(j;k)}_\ell \in \mathbb{R}^{577}$ and $\boldsymbol{u}^{(j;i)}_\ell \in \mathbb{R}^{577}$ satisfy the following requirements simultaneously;

- $\boldsymbol{u}^{(j;i)}_\ell$ and $\boldsymbol{u}^{(j;k)}_\ell$ are significantly correlated.
- $\boldsymbol{u}^{(j;k)}_\ell$ is expressed distinctly between healthy controls ($j \leq 8$) and patients ($j > 8$).
- $\boldsymbol{u}^{(j;i)}_\ell$ is expressed distinctly between healthy controls ($j \leq 8$) and patients ($j > 8$).

In order to validate these requirements visually, we show scatterplot for $1 \leq \ell \leq 9$ (Fig. 7.18). More or less all nine scatterplots look like satisfying the above requirements simultaneously. In order to select $\boldsymbol{u}_\ell$ and $\boldsymbol{v}_\ell$ used for miRNA and gene selection, respectively, we need to identify which ℓ satisfies the above requirements best. In order that, we propose several measures. First, we select miRNAs and genes. P-values are attributed as

$$P_k = P_{\chi^2}\left[> \left(\frac{u_{\ell k}}{\sigma_\ell} \right)^2 \right], \tag{7.55}$$

$$P_i = P_{\chi^2}\left[> \left(\frac{v_{\ell i}}{\sigma'_\ell} \right)^2 \right]. \tag{7.56}$$

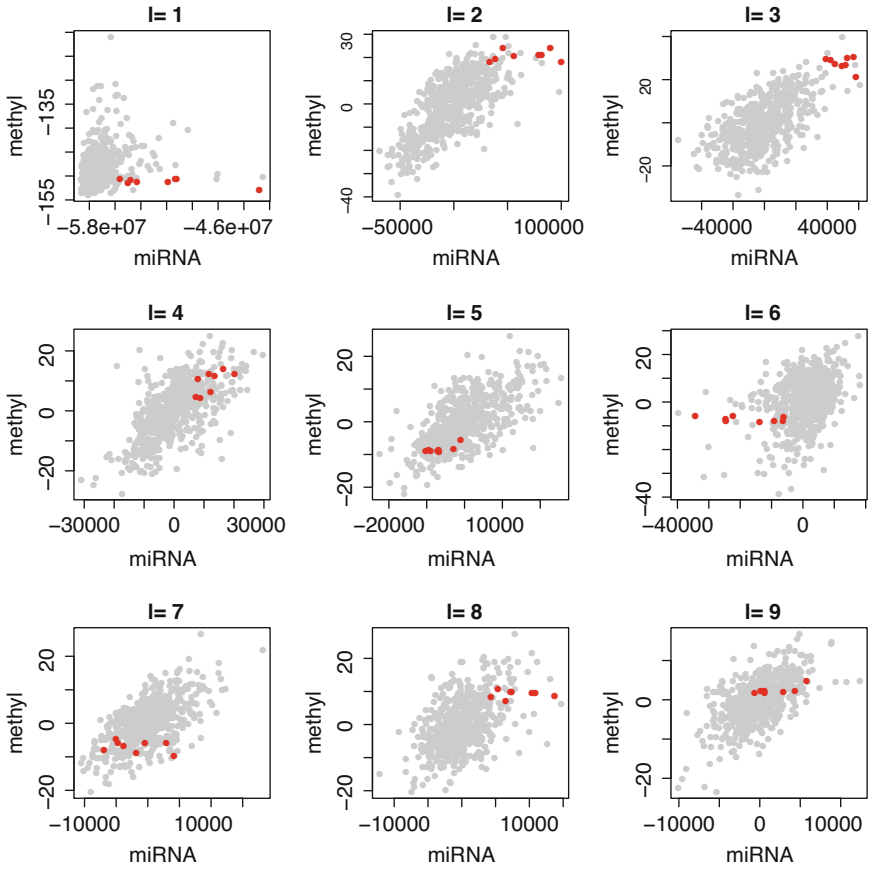

Fig. 7.18 Scatterplots of $\boldsymbol{u}_\ell^{(j;k)}$ (horizontal) and $\boldsymbol{u}_\ell^{(j;i)}$ (vertical) for $1 \leq \ell \leq 9$. Red filled circle: eight normal controls ($j \leq 8$), gray filled circles: ovarian cancer patients ($j > 8$)

P-values are adjusted by BH criterion and i and k associated with adjusted P-values less than 0.01 are selected. Then we require genes and miRNA selected similar to the above requirements as

- Selected genes and miRNAs are significantly correlated.
- Selected miRNAs are expressed distinctly between normal controls ($j \leq 8$) and patients ($j > 8$).
- Selected genes are methylated distinctly between normal controls ($j \leq 8$) and patients ($j > 8$).

In order that, we compute the followings:

(a) Correlation coefficient between $\boldsymbol{u}_\ell^{(j;i)}$ and $\boldsymbol{u}_\ell^{(j;k)}$.
(b) P-value attributed to the above correlation coefficients.

(c) P-values computed by t test that evaluates if $\boldsymbol{u}_\ell^{(j;k)}$ is distinct between normal control ($j \leq 8$) and patients ($j > 8$).
(d) P-values computed by t test that evaluates if $\boldsymbol{u}_\ell^{(j;i)}$ is distinct between normal control ($j \leq 8$) and patients ($j > 8$).
(e) Ratio of significantly correlated pairs of genes and miRNAs selected.
(f) Ratio of miRNA associated with adjusted P-values computed by t test that evaluates if selected miRNAs are expressed distinctly between normal control ($j \leq 8$) and patients ($j > 8$).
(g) Ratio of genes associated with adjusted P-values computed by t test that evaluates if selected genes are methylated distinctly between normal control ($j \leq 8$) and patients ($j > 8$).
(h) The number of selected miRNAs.
(i) The number of selected genes.

Here significant correlation is evaluated if associated BH criterion adjusted P-values are less than 0.01 (see page 112 for how to compute P-values attributed to correlation coefficients). Table 7.30 shows the result. $\ell = 3$ seems to be the best, because $\ell = 3$ is the best for the sixth and the seventh measures and the second best in the fifth measure; the fifth, sixth, and seventh measures are important because they are direct evaluations of selected genes and miRNAs. Because the number of selected genes and miRNAs do not vary depending on ℓ so much, it is the best to select $\ell = 3$. Because more than 88% of genes and miRNAs and their pairs satisfy the desired requirements in the above (88% is the smallest ratio (percentage) among requirements from (e) to (g) in Table 7.30), TD based unsupervised FE can be considered to have ability to select miRNAs and genes satisfying desired requirements mentioned above.

In order to see if other supervised methods can identify set of genes and miRNAs satisfying desired requirements, i.e., selected genes are methylated distinctly between healthy control and patients, miRNAs selected are expressed distinctly between healthy controls and patients, selected genes and miRNAs are significantly correlated, we apply t test to select genes methylated distinctly between healthy

Table 7.30 Measures that evaluate which ℓ satisfies the desired requirements best

ℓ	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
1	0.187	6.35×10^{-6}	6.25×10^{-3}	4.42×10^{-7}	–	**1.000**	–	2	0
2	**0.718**	$\mathbf{1.95 \times 10^{-92}}$	1.28×10^{-4}	1.21×10^{-11}	**0.944**	0.571	0.834	7	241
3	0.628	1.49×10^{-64}	$\mathbf{3.06 \times 10^{-8}}$	5.55×10^{-10}	0.884	**1.000**	**0.905**	7	284
4	0.649	2.45×10^{-70}	6.15×10^{-5}	1.02×10^{-4}	0.539	0.714	0.597	7	273
6	0.348	6.76×10^{-18}	1.68×10^{-3}	$\mathbf{5.71 \times 10^{-17}}$	0.350	0.375	0.674	8	132
7	0.624	1.27×10^{-63}	2.00×10^{-1}	7.65×10^{-7}	0.365	0.400	0.758	5	293
8	0.500	8.60×10^{-38}	1.33×10^{-4}	5.89×10^{-13}	0.274	0.833	0.775	6	231
9	0.593	3.50×10^{-56}	6.44×10^{-2}	3.35×10^{-5}	0.182	0.667	0.681	3	251

The number in the first row corresponds to the alphabetical list in the main text
Bold numbers are the best values within each category

controls and patients and miRNA expressed distinctly between healthy controls and patients. P-values are attributed to miRNAs and genes and adjusted by BH criterion. Then, 214 miRNAs and 19,395 genes associated with adjusted P-values less than 0.01 are selected. In order to see how much ratio of significantly correlated pairs among total $241 \times 19395 = 4{,}829{,}355$ pairs is, we compute correlation coefficients between them and attribute P-values to these pairs (see page 112 for how to compute P-values attributed to correlation coefficients). P-values are corrected by BH criterion and 555,391 pairs are associated with adjusted P-values less than 0.01. Because this is as small as 11.5% of 4829,355 pairs, t test is inferior to TD based unsupervised FE to identify genes and miRNAs satisfying desired requirements.

This poor performance might be because of the too many genes and miRNAs selected. P-values given by t test have strong tendency to reduce its value when many samples are available. In this example, because as many as 575 samples are available, even gene and miRNAs associated with small distinction are associated with small enough P-values. In order to avoid this difficulty, we reduce the number of genes and miRNAs selected by t test as many as those by TD based unsupervised FE, by selecting to ranked seven miRNA and 284 methylation probes attributed to genes based upon P-values computed by t test. Then among $7 \times 284 = 1967$ pairs, as small as 50 pairs are associated with adjusted P-values less than 0.01 attributed to correlation coefficient. Thus, only 2.5% of 1967 pairs are significantly correlated. Thus, the ratio decreases instead of increasing in opposed to the expectation.

It might be possible to select genes and miRNAs starting from identifying significantly correlated pairs before finding genes and miRNAs distinct between healthy control and patients. Then correlation coefficients are computed among all pairs of genes and miRNAs. P-values are attributed to correlation coefficient (see page 112 for how to compute P-values attributed to correlation coefficients) and are corrected by BH criterion. Then among $24{,}906 \times 723 = 18{,}007{,}038$ pairs, 1,197,772 pairs are associated with adjusted P-values less than 0.01. Unfortunately, these pairs include all genes and miRNAs. Thus, starting from pairs significantly correlated is not an effective strategy. This poor performance achieved by t test as well as correlation analysis demonstrates the difficulty of identifying gene and miRNAs satisfying desired requirement, i.e., selected genes are methylated distinctly between healthy control and patients, miRNAs selected are expressed distinctly between healthy controls and patients, selected genes and miRNAs are significantly correlated, which is easily achieved by TD based unsupervised FE.

Before closing this section, genes and miRNA selected should be biologically evaluated, too. First, 240 gene symbols associated with 284 probes are uploaded to DAVID (Table 7.31). At a glance, although it does not look deeply related to cancers, detailed investigation can alter this impression. This data is about ovarian cancer. The most major subtype is surface epithelial-stromal tumor which is known to be associated with keratinization [13]. Thus, the detection of keratinization as the most enriched term is reasonable, while the third enriched one is also related to keratinization. Because the fifth one, epidermis development, is the parent term of keratinization, it is also understandable.

Table 7.31 GO BP enrichment by the 274 gene symbols identified by TD based unsupervised FE for ovarian cancer data from TCGA

Category	Term	Genes count	%	P-value	Adjusted P-value
GOTERM_BP_DIRECT	Keratinization	14	6.2	9.3E−15	1.1E−11
GOTERM_BP_DIRECT	Peptide cross-linking	14	6.2	1.7E−14	9.6E−12
GOTERM_BP_DIRECT	Keratinocyte differentiation	15	6.6	2.8E−13	1.1E−10
GOTERM_BP_DIRECT	Acute-phase response	7	3.1	6.4E−6	1.8E−3
GOTERM_BP_DIRECT	Epidermis development	9	4.0	8.0E−6	1.8E−3

Adjusted P-values are by BH criterion

Next, the selected seven miRNAs are uploaded to DIANA-mirpath for the evaluation (Fig. 7.19). It is obvious that they are enriched with various cancers. Thus, the selected seven miRNAs are supposed to be related to cancers.

In conclusion, TD based unsupervised FE successfully identifies reasonable genes and miRNAs also from the biological point of view.

7.10 Summary

Because TD based unsupervised FE was more recently proposed than PCA based unsupervised FE, the examples of applications of TD based unsupervised FE introduced in this chapter are very limited. In spite of that, it still covers wide range of applications tried in the previous chapter using PCA based unsupervised FE: analysis of time course data set, integrated analysis of multi-omics data set, and identification of disease causing genes. In addition to this, it has new application target, e.g., application to in silico drug discovery.

The general procedure of application of TD based unsupervised FE is as follows. If there are no tensors available, generate case I or case II tensor of type I. Occasionally, it might be requires to generate type II tensor in order to reduce the required computational memory. If generated type II tensor is matrix, apply PCA. If not, apply HOSVD. If type II tensor is employed, generate missing singular value vectors by multiplying original tensor to obtained singular value vectors. Seek singular value vectors attributed to samples coincident with desired property, e.g., distinction between controls and treated samples. Then, in order to select singular value vectors attributed to features used for FE, core tensor is investigated. Singular value vectors that share core tensor with larger absolute values with singular value vectors attributed to samples associated with desired properties are selected. P-values are attributed to features using selected singular value vectors attributed to features with assuming χ^2 distributions. P-values are corrected by BH criterion and features associated with adjusted P-values less than 0.01.

This general procedure can be applied to wide range of bioinformatics topics depending upon what kind of singular value vectors attributed to samples are selected. In this sense, TD based unsupervised FE is expected to be applicable to wider range of biological problems other than those treated in this chapter.

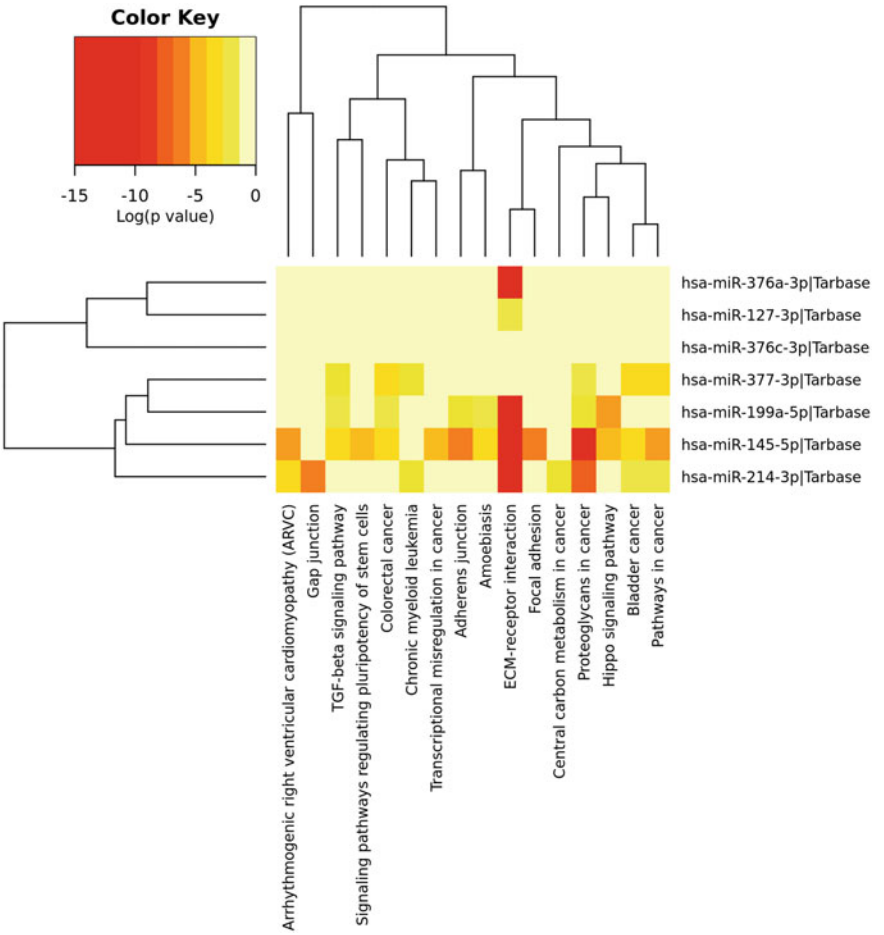

Fig. 7.19 Heatmap that summarize the results of DIANA-mirpath for the selected seven miRNAs, with specifying "pathways union" option

Appendix

Universarity of miRNA Transfection

Study 1

This data set includes transfection of three miRNAs, miR-200a, 200b, and 200c. The number of probes in microarray is as many as 43,376. For each of three, two paired experiments of treated and control samples. Treated and control sample measurement is performed by one microarray. Thus these two must be retrieved from it (columns annotated as gProcessedSignal and rProcessedSignal). Then, it

is possible to make a tensor, $x_{ij_1j_2j_3} \in \mathbb{R}^{43376\times3\times2\times2}$ where i stands for probes, j_1 stands for miRNAs, j_2 stands for two replicates, and j_3 stands for control vs treated samples. Nevertheless, it is not suitable for this specific case. If the number of components is two, automatically the two components of singular value vectors are $u_j = u_{j'}$ and $u_j = -u_{j'}$ where j and j' are each of two categories. The present purpose is to see if the components independent of category exist. This means, the setup that always results in the components independent of category is not good. Therefore, in this specific case, we format mRNA expression profiles as $x_{ij} \in \mathbb{R}^{43,376\times12}$ where $1 \le j \le 6$ and $7 \le j \le 12$ are control and treated samples, respectively. PCA is applied to x_{ij} such that PC score, $\boldsymbol{u}_\ell \in \mathbb{R}^{43376}$, and PC loading, $\boldsymbol{v}_\ell \in \mathbb{R}^{12}$, are attributed to probes and samples, respectively. As a result, we find that $\boldsymbol{v}_2$ represents distinct expression between control and treated samples, but independent of miRNAs transfected (Fig. 7.20). This suggests that there are non-negligible number of mRNAs affected by sequence-nonspecific off-target regulation. P-values are attributed to probes using the second PC score $\boldsymbol{u}_2$ with assuming χ^2 distribution as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{2i}}{\sigma_2}\right)^2 \right]. \tag{7.57}$$

P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

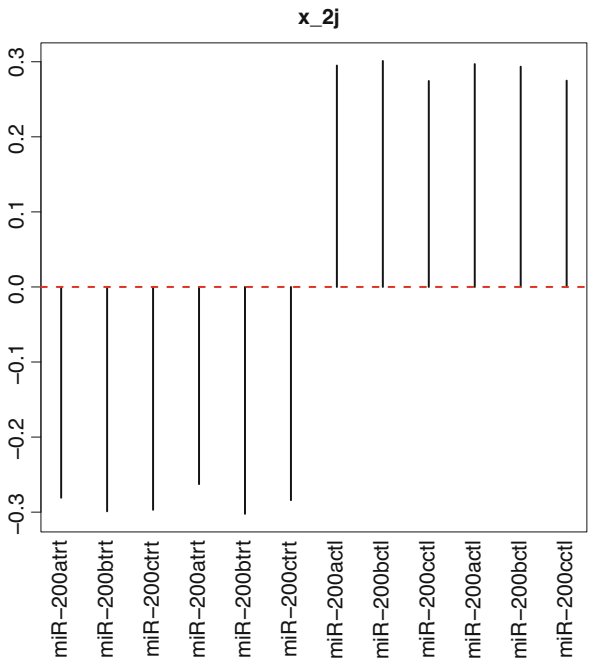

Fig. 7.20 The second PC loading, $\boldsymbol{v}_2$, obtained by PCA applied to x_{ij} made out of study 1

Study 2

This data set includes two miR-7 transfection experiments, two miR-128 transfection experiments, and three control experiments, normalized by mas5 procedure [17]. As mentioned in the beginning of the previous chapter, microarray technology measures photo emission of hybridized probes. Thus, various normalization procedures are applied. mas5 is one of such popular procedures, although I do not intend to explain mas5 in more detail, because it is beyond the scope of this textbook. Because of unmatched number of experiments of treated and control samples, they are difficult to be formatted in tensor. Thus it is instead formatted as matrix, $x_{ij} \in \mathbb{R}^{54675\times 7}$, where $j = 1, 2$ corresponds to miR-7 transfection $j = 3, 4$ corresponds to miR-128 transfection and $5 \leq j \leq 7$ correspond to control samples. PCA is applied to x_{ij} such that PC score, $\boldsymbol{u}_\ell \in \mathbb{R}^{54675}$, and PC loading, $\boldsymbol{v}_\ell \in \mathbb{R}^7$, are attributed to probes and samples, respectively. The result is a bit disappointing. In contrast to Fig. 7.20, we cannot find any PC loading that is constant independent of miRNAs transfected. Figure 7.21 shows the second PC loading, $\boldsymbol{v}_2$, which exhibits opposite signs between miR-7 transfection and miR-128 transfection. In spite of that, Fig. 7.21 still suggests the possibility of sequence-nonspecific off-target regulation. As mentioned previously, the only canonical function of miRNA is to downregulate target mRNAs. With only this function, it is impossible to assign opposite signs toward controls between miR-7 and miR-128 transfection as shown in Fig. 7.21. Downregulation can result in only same signs towards controls. At least, either of miR-7 or miR-128 transfection must be associated with sequence-nonspecific off-target regulation that can cause upregulation. Thus, we keep the selection of the second PC loading and assign P-values to probes as Eq. (7.57).

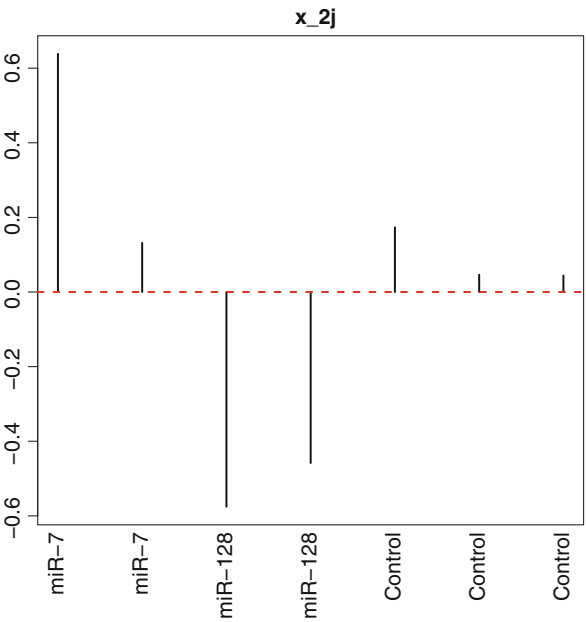

Fig. 7.21 The second PC loading, $\boldsymbol{v}_2$, obtained by PCA applied to x_{ij} made out of study 2

P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

Study 3

This data set includes two miR-7 transfection experiments, two miR-128 transfection experiments, and six control experiments, normalized by plier procedure [18]. Plier is yet another procedure that normalizes microarray, although I do not intend to explain plier in more detail, because it is beyond the scope of this textbook. Because number of experiments of treated and control samples, they are difficult to be formatted in tensor. Thus it is instead as matrix, $x_{ij} \in \mathbb{R}^{54675 \times 10}$, where $j = 1, 2$ corresponds to miR-7 transfection $j = 3, 4$ corresponds to miR-128 transfection and $5 \leq j \leq 10$ correspond to control samples. PCA is applied to x_{ij} such that PC score, $\boldsymbol{u}_\ell \in \mathbb{R}^{54675}$, and PC loading, $\boldsymbol{v}_\ell \in \mathbb{R}^{10}$, are attributed to probes and samples, respectively. The result is similar to study 2. In contrast to Fig. 7.20, we cannot find any PC loading that is constant independent of miRNAs transfected. Figure 7.22 shows the second PC loading, $\boldsymbol{v}_2$, which exhibits opposite signs between miR-7 transfection and miR-128 transfection. As in the study 2, we keep the selection of the second PC loading and assign P-values to probes as Eq. (7.57). P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

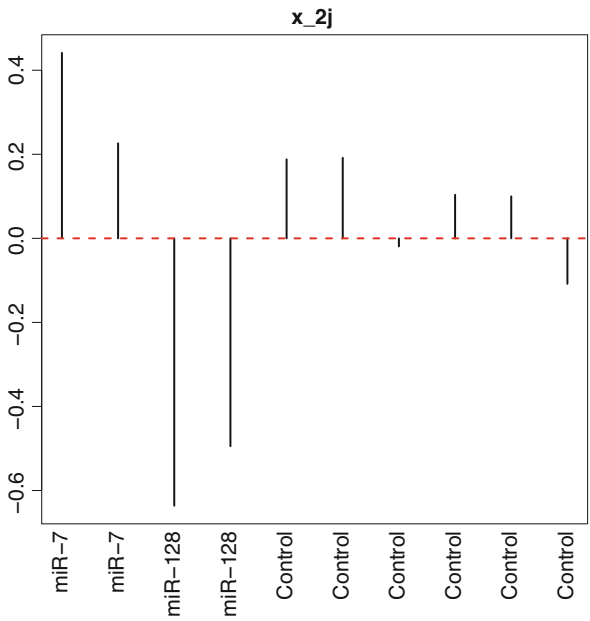

Fig. 7.22 The second PC loading, $\boldsymbol{v}_2$, obtained by PCA applied to x_{ij} made out of study 3

Study 4

This data set includes two replicates of nine transfected miRNAs (miR-7/9/122a/128a/132/133a/142/148b/181a) and corresponding 18 control samples. Thus, the total number of samples is 36. This is successfully formatted as tensor, $x_{ijk} \in \mathbb{R}^{23651\times 18\times 2}$ where i stands for probes, j stands for nine miRNAs transfection times two biological replicates, and k is control and treated samples. We apply HOSVD algorithm, Fig. 3.8, to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{23651} \sum_{\ell_2=1}^{18} \sum_{\ell_3=1}^{2} G(\ell_1, \ell_2, \ell_3) u^{(i)}_{\ell_1 i} u^{(j)}_{\ell_2 j} u^{(k)}_{\ell_3 k} \tag{7.58}$$

where $\boldsymbol{u}^{(i)}_{\ell_1} \in \mathbb{R}^{23651}, \boldsymbol{u}^{(j)}_{\ell_2} \in \mathbb{R}^{18}, \boldsymbol{u}^{(k)}_{\ell_3} \in \mathbb{R}^{2}$ are singular value vectors and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{23651\times 18\times 2}$ is a core tensor. Now we need to find $\boldsymbol{u}^{(k)}_{\ell_3}$ satisfying $u^{(k)}_{\ell_3 1} = -u^{(k)}_{\ell_3 2}$; $\ell_3 = 2$ turns out to satisfy this requirement. On the other hand, we need to find $\boldsymbol{u}^{(j)}_{\ell_2}$ satisfying $u^{(j)}_{\ell_2 j}$ = constant; $\ell_2 = 1$ turns out to satisfy this requirement (Fig. 7.23). After investigating which $G(\ell_1, 1, 2)$ has the largest absolute value, we find that $\ell_1 = 6$. P-values are attributed to probes using the sixth PC score $\boldsymbol{u}^{(i)}_6$ with assuming χ^2 distribution as

$$P_i = P_{\chi^2}\left[> \left(\frac{u^{(i)}_{6i}}{\sigma_6} \right)^2 \right]. \tag{7.59}$$

P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

Study 5

This data set includes four profiles to which mock and cel-miR-67 miR-509/199a-3p are transfected. We format it to matrix $x_{ij} \in \mathbb{R}^{41539\times 4}$. PCA is applied to x_{ij} and the second PC loading, $\boldsymbol{v}_2$, is selected as that exhibits distinction between mock + cel-miR-67 and miR-509/199a-3p (Fig. 7.24). Although outcome cannot be said very promising, because $\boldsymbol{v}_2$ is best fitted with the requirement, P-values are attributed to probes using Eq. (7.57). P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

Study 6

This data set includes transfection of eight miRNAs, miR-10a-5p, 150-3p/5p, 148a-3p/5p, 499a-5p, 455-3p. The number of probes in microarray is as many as 62,976. The number of samples is 16 composed of combination of miRNAs and cell lines.

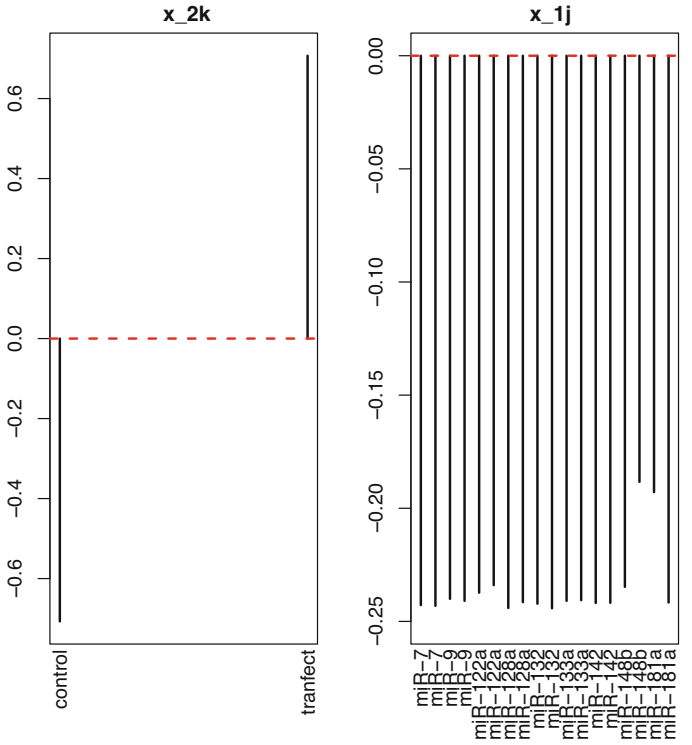

Fig. 7.23 The second singular value vector, $\boldsymbol{u}_2^{(k)}$, attributed to control and treated samples, and the first singular value vector, $\boldsymbol{u}_1^{(j)}$, attributed to miRNAs and replicates, obtained by HOSVD applied to x_{ijk} made out of study 4

Not all miRNAs are used equally. For each of 16, two paired experiments of treated and control samples. Treated and control sample measurement is performed by one microarray. Thus these two must be retrieved from it (columns annotated as gProcessedSignal and rProcessedSignal). This is successfully formatted as tensor, $x_{ijk} \in \mathbb{R}^{62976\times 16\times 2}$ where i stands for probes, j stands for combinations of eight miRNAs transfection and cell lines, and k is control and treated samples. We apply HOSVD algorithm, Fig. 3.8, to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{62976} \sum_{\ell_2=1}^{16} \sum_{\ell_3=1}^{2} G(\ell_1, \ell_2, \ell_3) u_{\ell_1 i}^{(i)} u_{\ell_2 j}^{(j)} u_{\ell_3 k}^{(k)} \tag{7.60}$$

where $\boldsymbol{u}_{\ell_1}^{(i)} \in \mathbb{R}^{62976}, \boldsymbol{u}_{\ell_2}^{(j)} \in \mathbb{R}^{16}, \boldsymbol{u}_{\ell_3}^{(k)} \in \mathbb{R}^2$ are singular value vectors and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{62976\times 16\times 2}$ is a core tensor. Now we need to find $\boldsymbol{u}_{\ell_3}^{(k)}$ satisfying

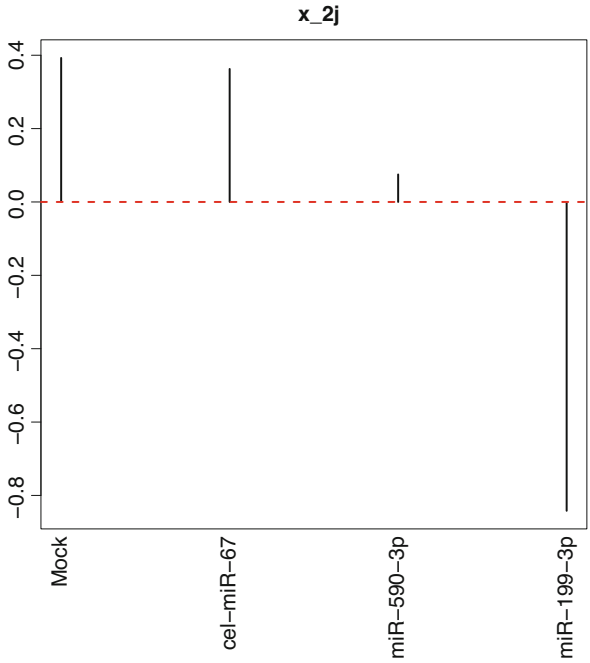

Fig. 7.24 The second PC loading, $\boldsymbol{v}_2$, obtained by PCA applied to x_{ij} made out of study 5

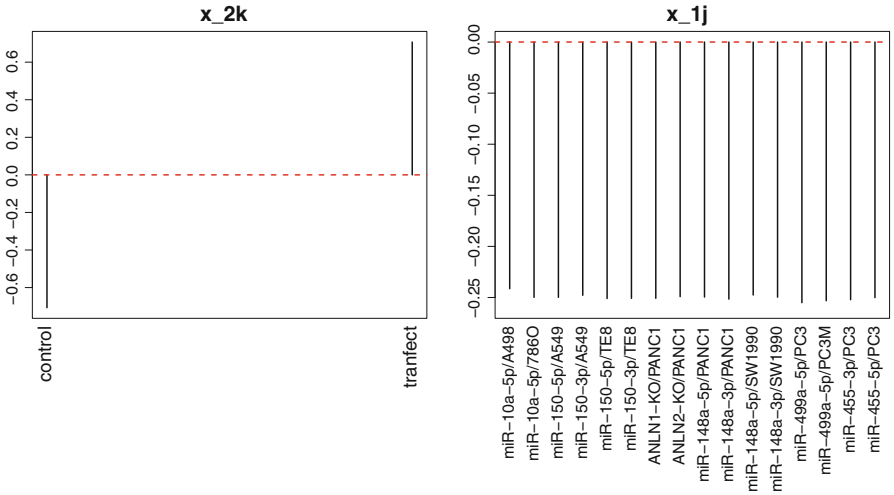

Fig. 7.25 The second singular value vector, $\boldsymbol{u}_2^{(k)}$, attributed to the various combinations of control and cell lines, and the first singular value vector, $\boldsymbol{u}_1^{(j)}$, attributed to miRNAs and replicates, obtained by HOSVD applied to x_{ijk} made out of study 6

$u_{\ell_3 1}^{(k)} = -u_{\ell_3 2}^{(k)}$; $\ell_3 = 2$ turns out to satisfy this requirement. On the other hand, we need to find $\boldsymbol{u}_{\ell_2}^{(j)}$ satisfying $u_{\ell_2 j}^{(j)}$ = constant; $\ell_2 = 1$ turns out to satisfy this requirement (Fig. 7.25). After investigating which $G(\ell_1, 1, 2)$ has the largest

absolute value, we find that $\ell_1 = 7$. P-values are attributed to probes using the seventh PC score $\boldsymbol{u}_7^{(i)}$ with assuming χ^2 distribution as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{7i}^{(i)}}{\sigma_7} \right)^2 \right]. \tag{7.61}$$

P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

Study 7

This data set includes transfection of nine miR-205/29a/144-3p/5p, 210, 23b, 221/222/223. The number of probes in microarray is as many as 62,976. The number of samples is 19 composed of combination of miRNAs and cell lines. Not all miRNAs are used equally. For each of 19, two paired experiments of treated and control samples. Treated and control sample measurement is performed by one microarray. Thus these two must be retrieved from it (columns annotated as gProcessedSignal and rProcessedSignal). This is successfully formatted as tensor, $x_{ijk} \in \mathbb{R}^{62976\times 19\times 2}$ where i stands for probes, j stands for combinations of eight miRNAs transfection and cell lines, and k is control and treated samples. We apply HOSVD algorithm, Fig. 3.8, to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{62976} \sum_{\ell_2=1}^{19} \sum_{\ell_3=1}^{2} G(\ell_1, \ell_2, \ell_3) u_{\ell_1 i}^{(i)} u_{\ell_2 j}^{(j)} u_{\ell_3 k}^{(k)} \tag{7.62}$$

where $\boldsymbol{u}_{\ell_1}^{(i)} \in \mathbb{R}^{62976}, \boldsymbol{u}_{\ell_2}^{(j)} \in \mathbb{R}^{19}, \boldsymbol{u}_{\ell_3}^{(k)} \in \mathbb{R}^2$ are singular value vectors and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{62976\times 19\times 2}$ is a core tensor. Now we need to find $\boldsymbol{u}_{\ell_3}^{(k)}$ satisfying $u_{\ell_3 1}^{(k)} = -u_{\ell_3 2}^{(k)}$; $\ell_3 = 2$ turns out to satisfy this requirement. On the other hand, we need to find $\boldsymbol{u}_{\ell_2}^{(j)}$ satisfying $u_{\ell_2 j}^{(j)}$ = constant; $\ell_2 = 1$ turns out to satisfy this requirement (Fig. 7.26). After investigating which $G(\ell_1, 1, 2)$ has the larger absolute values, we find that $\ell_1 = 2, 3$. P-values are attributed to probes using the second and third PC scores $\boldsymbol{u}_{\ell_1}^{(i)}$, $\ell_1 = 2, 3$ with assuming χ^2 distribution as

$$P_i = P_{\chi^2}\left[> \sum_{\ell_1=2}^{3} \left(\frac{u_{\ell_1 i}^{(i)}}{\sigma_{\ell_1}} \right)^2 \right]. \tag{7.63}$$

P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

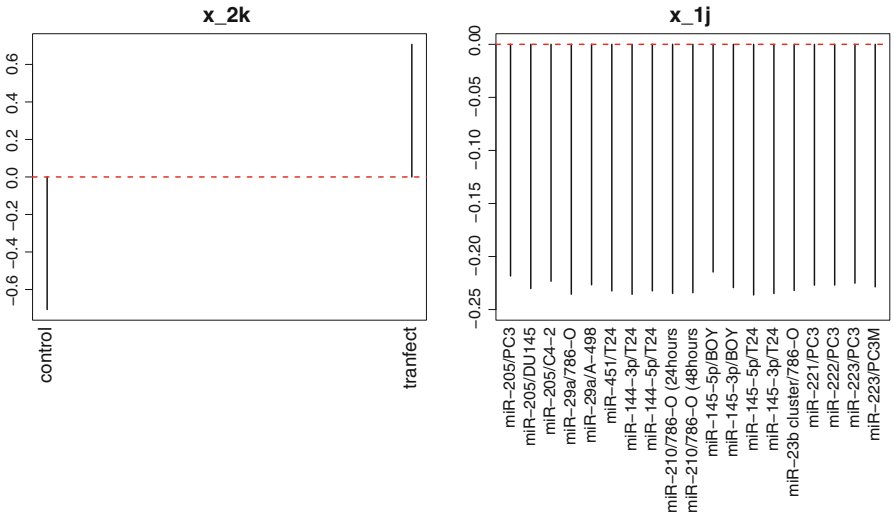

Fig. 7.26 The second singular value vector, $\boldsymbol{u}_2^{(k)}$, attributed to control and treated samples, and the first singular value vector, $\boldsymbol{u}_1^{(j)}$, attributed to the combinations of miRNAs and cell lines, obtained by HOSVD applied to x_{ijk} made out of study 7

Study 8

This data set includes transfection of two miRNAs, miR-146a/b. The number of probes in microarray is as many as 43,379. The number of samples is 18 composed of six miR-146a OE, four miR-146b OE, and eight miR-146a KO. For each of 19, two paired experiments of treated and control samples. Treated and control sample measurement is performed by one microarray. Thus these two must be retrieved from it (columns annotated as gProcessedSignal and rProcessedSignal). This is successfully formatted as tensor, $x_{ijk} \in \mathbb{R}^{43379\times 18\times 2}$ where i stands for probes, j stands for combinations of eight miRNAs transfection and cell lines, and k is control and treated samples. We apply HOSVD algorithm, Fig. 3.8, to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{43379} \sum_{\ell_2=1}^{18} \sum_{\ell_3=1}^{2} G(\ell_1, \ell_2, \ell_3) u_{\ell_1 i}^{(i)} u_{\ell_2 j}^{(j)} u_{\ell_3 k}^{(k)} \tag{7.64}$$

where $\boldsymbol{u}_{\ell_1}^{(i)} \in \mathbb{R}^{43379}, \boldsymbol{u}_{\ell_2}^{(j)} \in \mathbb{R}^{18}, \boldsymbol{u}_{\ell_3}^{(k)} \in \mathbb{R}^2$ are singular value vectors and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{43379\times 18\times 2}$ is a core tensor. Now we need to find $\boldsymbol{u}_{\ell_3}^{(k)}$ satisfying $u_{\ell_3 1}^{(k)} = -u_{\ell_3 2}^{(k)}$; $\ell_3 = 2$ turns out to satisfy this requirement. On the other hand, we need to find $\boldsymbol{u}_{\ell_2}^{(j)}$ satisfying $u_{\ell_2 j}^{(j)}$ = constant; $\ell_2 = 1$ turns out to satisfy this requirement (Fig. 7.27). After investigating which $G(\ell_1, 1, 2)$ has the largest absolute value, we find that $\ell_1 = 5$. P-values are attributed to probes using the fifth

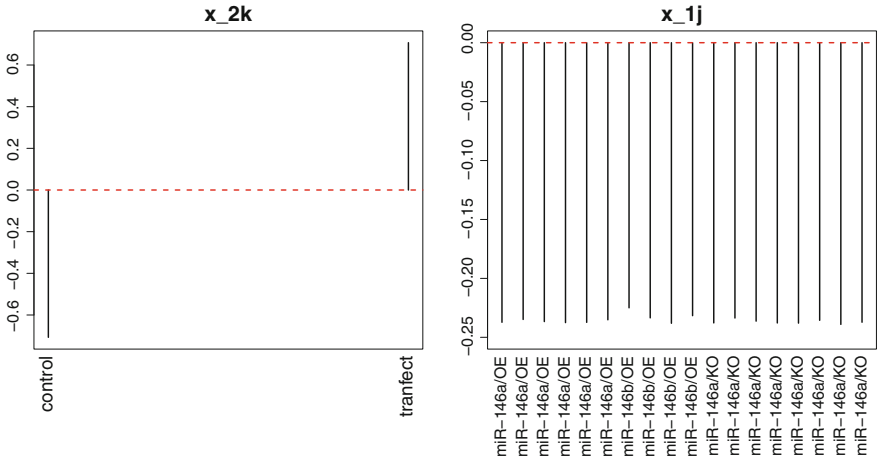

Fig. 7.27 The second singular value vector, $\boldsymbol{u}_2^{(k)}$, attributed to control and treated samples, and the first singular value vector, $\boldsymbol{u}_1^{(j)}$, attributed to miRNAs and replicates, obtained by HOSVD applied to x_{ijk} made out of study 8

PC score $\boldsymbol{u}_5^{(i)}$ with assuming χ^2 distribution as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{5i}^{(i)}}{\sigma_5} \right)^2 \right]. \tag{7.65}$$

P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

Study 9

This data set includes transfection of two miRNAs, miR-107/181b. transfected to HeLa cell lines. The number of probes in microarray is as many as 9987. The number of samples is 18 composed of six controls, two anti-miR-107, four miR-107, two anti-miR-181b, and four miR-181b transfected samples. This is successfully formatted as tensor, $x_{ijk} \in \mathbb{R}^{9987\times 16\times 3}$ where i stands for probes, j stands for replicates, and k is control, miR-107 and miR-181b. We apply HOSVD algorithm, Fig. 3.8, to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{9987} \sum_{\ell_2=1}^{6} \sum_{\ell_3=1}^{3} G(\ell_1, \ell_2, \ell_3) u_{\ell_1 i}^{(i)} u_{\ell_2 j}^{(j)} u_{\ell_3 k}^{(k)} \tag{7.66}$$

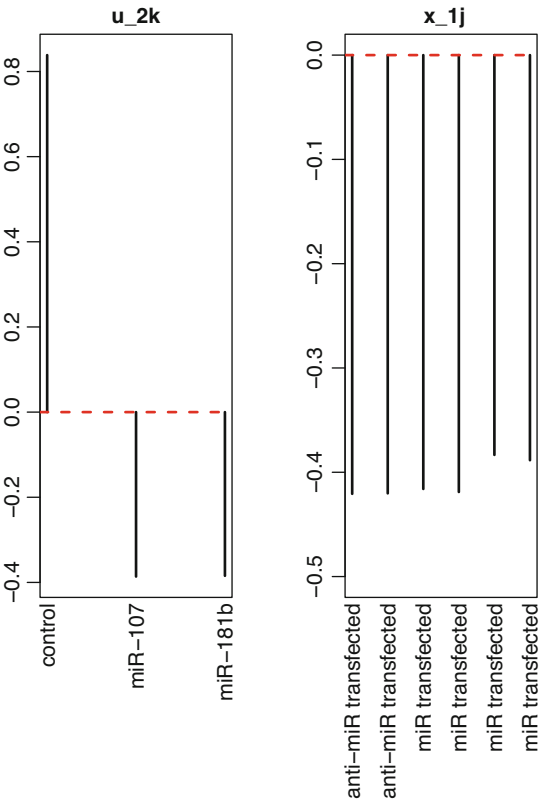

Fig. 7.28 The second singular value vector, $\boldsymbol{u}_2^{(k)}$, attributed to control, miR-107 and miR-181b transfection, and the first singular value vector, $\boldsymbol{u}_1^{(j)}$, attributed to miRNAs and replicates, obtained by HOSVD applied to x_{ijk} made out of study 9

where $\boldsymbol{u}_{\ell_1}^{(i)} \in \mathbb{R}^{9987}, \boldsymbol{u}_{\ell_2}^{(j)} \in \mathbb{R}^6, \boldsymbol{u}_{\ell_3}^{(k)} \in \mathbb{R}^3$ are singular value vectors and $G(\ell_1, \ell_2, \ell_3) \in \mathbb{R}^{9987\times 6\times 3}$ is a core tensor. Now we need to find $\boldsymbol{u}_{\ell_3}^{(k)}$ satisfying $u_{\ell_3 1}^{(k)} = -u_{\ell_3 2}^{(k)} = -u_{\ell_3 3}^{(k)}$; $\ell_3 = 2$ turns out to satisfy this requirement. On the other hand, we need to find $\boldsymbol{u}_{\ell_2}^{(j)}$ satisfying $u_{\ell_2 j}^{(j)}$ = constant; $\ell_2 = 1$ turns out to satisfy this requirement (Fig. 7.28). After investigating which $G(\ell_1, 1, 2)$ has the largest absolute value, we find that $\ell_1 = 2$. P-values are attributed to probes using the second PC score $\boldsymbol{u}_2^{(i)}$ with assuming χ^2 distribution as Eq. (7.57). P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

Study 10

Everything is the same as study nine other than that transfected cell line is HEK 293 cell line (see Fig. 7.29 for singular value vectors selected).

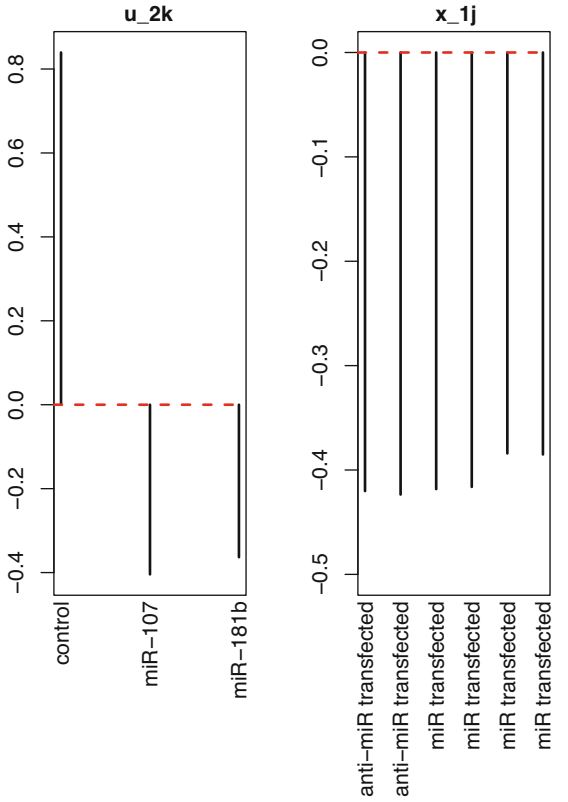

Fig. 7.29 The second singular value vector, $\boldsymbol{u}_2^{(k)}$, attributed to control, miR-107 and miR-181b transfection, and the first singular value vector, $\boldsymbol{u}_1^{(j)}$, attributed to miRNAs and replicates, obtained by HOSVD applied to x_{ijk} made out of study 10

Study 11

This data set includes transfection of a miRNA, miR-181b transfected to SH-SY5Y cell line. The number of probes in microarray is as many as 9987. The number of samples is eight composed of four controls, two anti-miR-181b, and two miR-181b transfected samples. This is successfully formatted as tensor, $x_{ijk} \in \mathbb{R}^{9987\times 4\times 2}$ where i stands for probes, j stands for replicates, and k is control and miR-181b. We apply HOSVD algorithm, Fig. 3.8, to x_{ijk} as

$$x_{ijk} = \sum_{\ell_1=1}^{9987}\sum_{\ell_2=1}^{4}\sum_{\ell_3=1}^{2} G(\ell_1,\ell_2,\ell_3)u_{\ell_1 i}^{(i)}u_{\ell_2 j}^{(j)}u_{\ell_3 k}^{(k)} \tag{7.67}$$

where $\boldsymbol{u}_{\ell_1}^{(i)} \in \mathbb{R}^{9987}, \boldsymbol{u}_{\ell_2}^{(j)} \in \mathbb{R}^4, \boldsymbol{u}_{\ell_3}^{(k)} \in \mathbb{R}^2$ are singular value vectors and $G(\ell_1,\ell_2,\ell_3) \in \mathbb{R}^{9987\times 4\times 2}$ is a core tensor. Now we need to find $\boldsymbol{u}_{\ell_3}^{(k)}$ satisfying $u_{\ell_3 1}^{(k)} = -u_{\ell_3 2}^{(k)}$; $\ell_3 = 2$ turns out to satisfy this requirement. On the other hand, we need to find $\boldsymbol{u}_{\ell_2}^{(j)}$ satisfying $u_{\ell_2 j}^{(j)} = \text{constant}$; $\ell_2 = 1$ turns out to satisfy

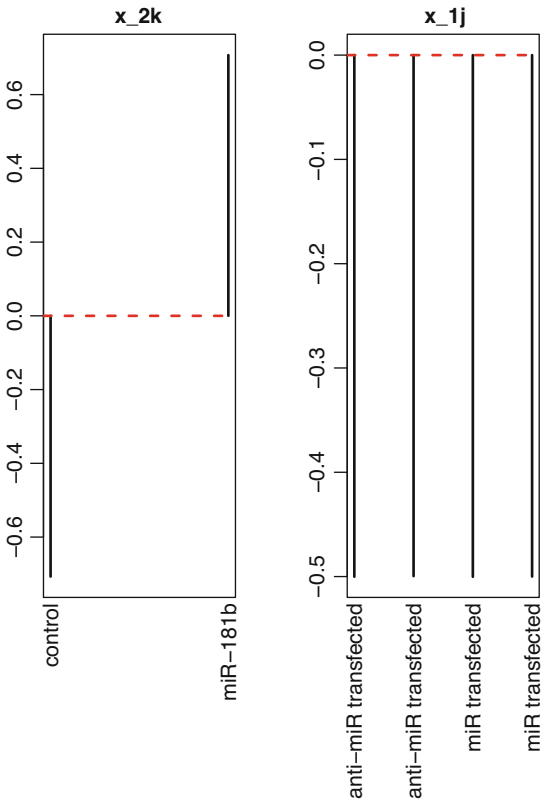

Fig. 7.30 The second singular value vector, $\boldsymbol{u}_2^{(k)}$, attributed to control and miR-181b transfection, and the first singular value vector, $\boldsymbol{u}_1^{(j)}$, attributed to miRNAs and replicates, obtained by HOSVD applied to x_{ijk} made out of study 11

this requirement (Fig. 7.30). After investigating which $G(\ell_1, 1, 2)$ has the largest absolute value, we find that $\ell_1 = 2$. P-values are attributed to probes using the second PC score $\boldsymbol{u}_2^{(i)}$ with assuming χ^2 distribution as Eq. (7.57). P-values are corrected by BH criterion and probes associated with adjusted P-values less than 0.01 are selected.

Drug Discovery From Gene Expression: II

Heart Failure

Human gene expression profiles are downloaded from GEO with GEO ID 57345. File used is GSE57345-GPL11532_series_matrix.txt.gz. Rat heart gene expression profiles are downloaded from GEO with GEO ID GSE59905. Files used are GSE59905-GPL5426_series_matrix.txt.gz, and GSE59905-GPL5425_series_matrix.txt.gz. 3937 genes are shared between human and rat. Case II tensor, $x_{ij_1j_2j_3}$, is generated as

$$x_{ij_1j_2j_3} = x_{ij_1j_2}x_{ij_3}. \tag{7.68}$$

HOSVD algorithm, Fig. 3.8, is applied to $x_{ij_1j_2j_3}$.

At first, we try to find $\boldsymbol{u}_{\ell_3}^{(j_3)}$ associated with significant distinction between three classes, healthy control, idiopathic dilated cardiomyopathy, ischemic stroke, by applying categorical regression

$$u_{\ell_3 j_3}^{(j_3)} = a_{\ell_3} + \sum_{s=1}^{3} b_{\ell_3 s}\delta_{j_3 s} \tag{7.69}$$

P-values computed by categorical regression are corrected by BH criterion. Then we found that $\ell_3 = 2, 3, 5, 17, 313$ are associated with adjusted P-values less than 0.01, raw P-values of which are 1.65×10^{-17}, 1.00×10^{-39}, 1.29×10^{-4}, 4.97×10^{-6} and 1.554×10^{-4}. Among them we select $\ell_3 = 2, 3$ because they have more contribution than others. Figure 7.31a shows the $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $1 \le \ell_3 \le 3$.

Next we try to identify $\boldsymbol{u}_{\ell_2}^{(j_2)}$ associated with significant time dependence. Figure 7.31b shows the $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$. The correlation coefficients between $\boldsymbol{u}_{\ell_2}^{(j_2)}$ and (1/4,1,3,5) are -0.72, -0.82, 0.51, and -0.09. Then $\ell_2 = 2$ with largest absolute value is selected. Then we need to find $\boldsymbol{u}_{\ell_1}^{(j_1)}$ and $\boldsymbol{u}_{\ell_4}^{(i)}$ associated with larger absolute $G(\ell_1, 2, 2, \ell_4)$ or $G(\ell_1, 2, 3, \ell_4)$ in order to select compounds j_1 and genes i associated with time dependence and distinction between patients and healthy controls simultaneously. In order that, we list top 20 $G(\ell_1, 2, 2, \ell_4)$ or $G(\ell_1, 2, 3, \ell_4)$ (Table 7.32). Because G gradually decreases, we cannot select specific cut off. Thus, tentatively, we select ℓ_1 and ℓ_4 associated with top 10 Gs; $\ell_1 = 2$ and $\ell_4 = 21, 25, 27, 28, 33, 36, 37, 38, 41, 42$. Figure 7.31c shows $\boldsymbol{u}_2^{(j_1)}$. Forty three outlier drugs, $\left|u_{2j_1}^{(j_1)}\right| > 0.1$, blue parts, are selected, by visual inspection, because P-values computed from $\boldsymbol{u}_2^{(j_1)}$ and corrected by BH criterion cannot be less than 0.01. On the other hand, P-values are attributed to ith gene as

$$P_i = P_{\chi^2}\left[> \sum_{\ell_4 = 21,25,27,28,33,36,37,38,41,42} \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}}\right)^2 \right] \tag{7.70}$$

P-values are corrected by BH criterion and 274 genes associated with adjusted P-values less than 0.01 are selected.

PTSD

PTSD model rat amygdala and hippocampus gene expression are downloaded from GEO with GEO ID GSE60304. A file GSE60304_series_matrix.txt.gz is used. Gene expression profiles of the brain for drug treatments of rats are

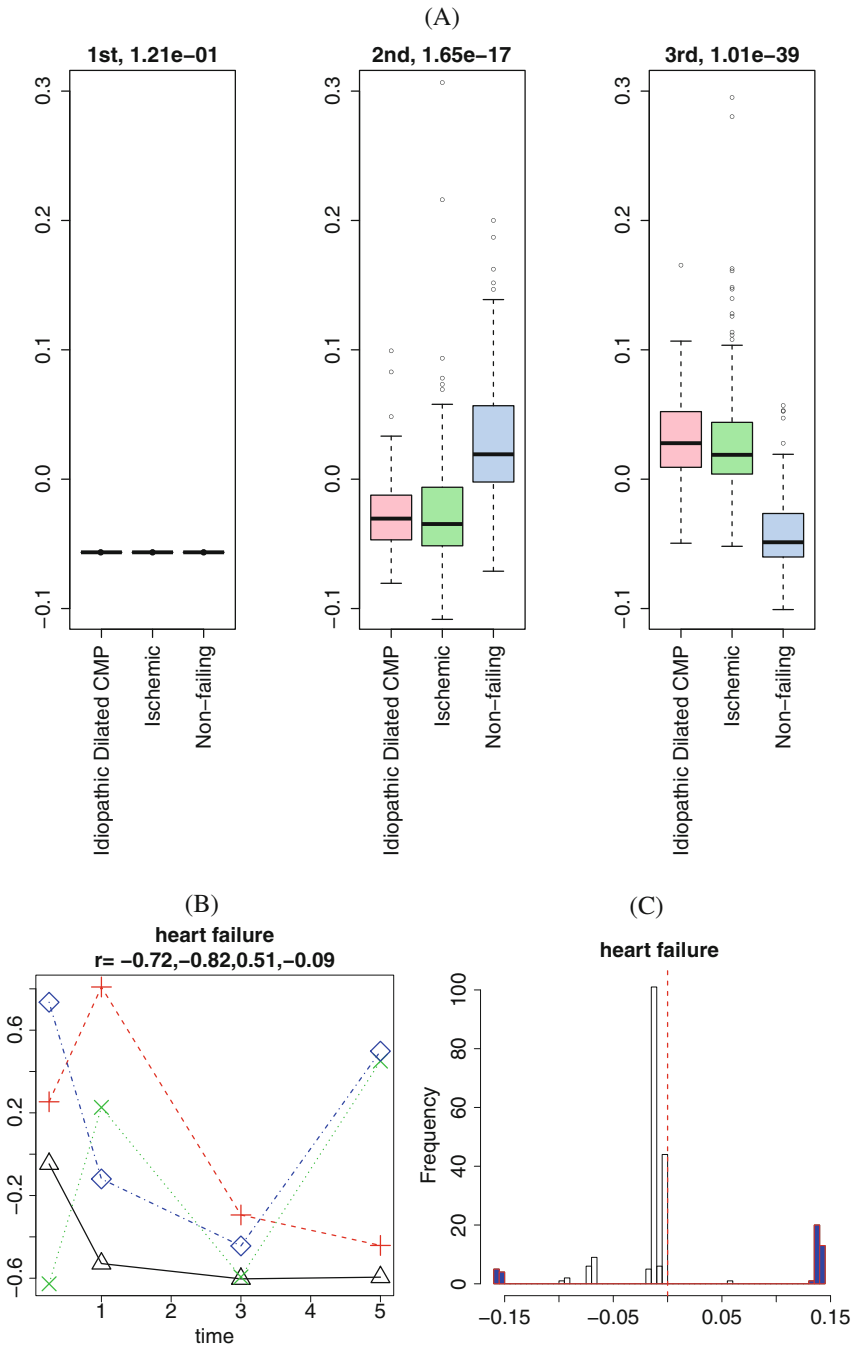

Fig. 7.31 (**a**) $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $1 \le \ell_3 \le 3$, P-values are computed by categorical regression, Eq. (7.69). (**b**) $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$, open triangle: $\ell_2 = 1$, red plus symbol: $\ell_2 = 2$, green cross symbol: $\ell_2 = 3$, blue diamond: $\ell_2 = 4$. r: correlation coefficient. (**c**) Histogram of $\boldsymbol{u}_2^{(j_1)}$. Blue parts are selected ones. Vertical red broken line is 0

Table 7.32 Top 20 $G(\ell_1, 2, 2, \ell_4)$ or $G(\ell_1, 2, 3, \ell_4)$

Rank	1	2	3	4	5	6	7	8	9	10
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	2	3	2	2	3	3	2	3	3	2
ℓ_4	27	38	33	28	41	37	21	36	42	25
$G(\ell_1, 2, \ell_3, \ell_4)$	66.2	−43.7	40.7	−40.2	38.2	−31.6	28.5	−26.8	−26.2	−26.2

Rank	11	12	13	14	15	16	17	18	19	20
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	2	3	2	2	3	3	2	2	2	3
ℓ_4	40	29	31	39	32	33	26	11	18	31
$G(\ell_1, 2, \ell_3, \ell_4)$	−25.5	25.2	−22.6	21.8	20.7	−19.7	−19.5	−18.2	−17.3	15.4

downloaded from GEO with GEO ID GSE59895. Files used are GSE59895-GPL5425_series_matrix.txt.gz and GSE59895-GPL5426_series_matrix.txt.gz. Case II tensor, $x_{ij_1j_2j_3j_4j_5}$, is generated as

$$x_{ij_1j_2j_3j_4j_5} = x_{ij_1j_2}x_{ij_3j_4}x_{ij_3j_5}. \tag{7.71}$$

HOSVD algorithm, Fig. 3.8, is applied to $x_{ij_1j_2j_3j_4j_5}$.

In order to identify $\boldsymbol{u}_{\ell_4}^{(j_4)}$ and $\boldsymbol{u}_{\ell_5}^{(j_5)}$ associated with three classes, control samples, minimal behavioral response samples, and extreme behavioral response samples, by applying categorical regression,

$$u_{\ell j_4}^{(j_4)} = a_\ell + \sum_{s=1}^{3} b_{\ell s}\delta_{j_4 s} \tag{7.72}$$

$$u_{\ell j_5}^{(j_5)} = a_\ell + \sum_{s=1}^{3} b_{\ell s}\delta_{j_5 s} \tag{7.73}$$

where regression coefficients are shared between $\ell_4 = \ell_5 = \ell$. P-values computed by categorical regression are corrected by BH criterion. Then, only $\ell = 3$ is associated with adjusted P-values less than 0.05 (Fig. 7.32a).

Next we try to identify $\boldsymbol{u}_{\ell_2}^{(j_2)}$ associated with significant time dependence. Figure 7.32b shows the $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$. The correlation coefficients between $\boldsymbol{u}_{\ell_2}^{(j_2)}$ and (1/4,1,3,5) are $-0.75, -0.81, -0.30$, and 0.50. Then $\ell_2 = 2$ with largest absolute value is selected. Then we need to find $\boldsymbol{u}_{\ell_1}^{(j_1)}$ and $\boldsymbol{u}_{\ell_6}^{(i)}$ associated with larger absolute $G(\ell_1, 2, \ell_3, 3, 3, \ell_6)$ in order to select compounds j_1 and genes i associated with time dependence and distinction between patients and healthy controls simultaneously. In order that, we list top 20 $G(\ell_1, 2, \ell_3, 3, 3, \ell_6)$ (Table 7.33). Because G gradually decreases, we cannot select specific cut off. Thus, tentatively, we select ℓ_1 and ℓ_4 associated with top 10 Gs; $\ell_1 = 2$ and

Table 7.33 Top 20 $G(\ell_1, 2, \ell_3, 3, 3, \ell_6)$

Rank	1	2	3	4	5	6	7	8	9	10
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	1	1	1	1	1	1	1	1	1	1
ℓ_6	81	84	88	77	85	75	83	90	90	102
$G(\ell_1, 2, \ell_3, 3, 3, \ell_6)$	−0.133	0.112	0.110	−0.078	0.075	−0.075	0.074	0.069	0.069	−0.063

Rank	11	12	13	14	15	16	17	18	19	20
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	1	1	2	2	2	1	2	2	1	2
ℓ_6	76	80	94	76	128	285	86	286	92	282
$G(\ell_1, 2, \ell_3, 3, 3, \ell_6)$	−0.063	0.062	0.054	−0.054	−0.053	−0.052	0.048	0.047	0.045	0.045

$\ell_6 = 75, 77, 81, 83, 84, 85, 88, 89, 90, 102$. Figure 7.32c shows $\boldsymbol{u}_2^{(j_1)}$. Six outlier drugs, $u_{2j_1}^{(j_1)} < -0.2$ and $u_{1j_1}^{(j_1)} < -0.15$, blue parts, are selected, by visual inspection, because P-values computed from $\boldsymbol{u}_2^{(j_1)}$ and corrected by BH criterion cannot be less than 0.01. On the other hand, P-values are attributed to ith gene as

$$P_i = P_{\chi^2}\left[> \sum_{\ell_6=75,77,81,83,84,85,88,89,90,102} \left(\frac{u_{\ell_6 i}}{\sigma_{\ell_6}} \right)^2 \right] \tag{7.74}$$

P-values are corrected by BH criterion and 374 genes associated with adjusted P-values less than 0.01 are selected.

ALL

Bone marrow gene expression profiles of drug treated rats are downloaded from GEO with GEO ID GSE59894, and ALL human bone marrow gene expression is from GEO with GEO ID GSE67684. Used files are GSE67684-GPL570_series_matrix.txt.gz, GSE67684-GPL96_series_matrix.txt.gz, GSE59894-GPL5425_series_matrix.txt.gz, and GSE59894-GPL5426_series_matrix.txt.gz. In this case both gene expression profiles are time dependent. ALL human bone marrow gene expression profiles are measured at four times points, 0, 8, 15, and 33 days after a remission induction therapy. Case II tensor, $x_{ij_1j_2j_3j_4}$ is obtained as

$$x_{ij_1j_2j_3j_4} = x_{ij_1j_2}x_{ij_3j_4} \tag{7.75}$$

HOSVD algorithm, Fig. 3.8, is applied to $x_{ij_1j_2j_3j_4}$.

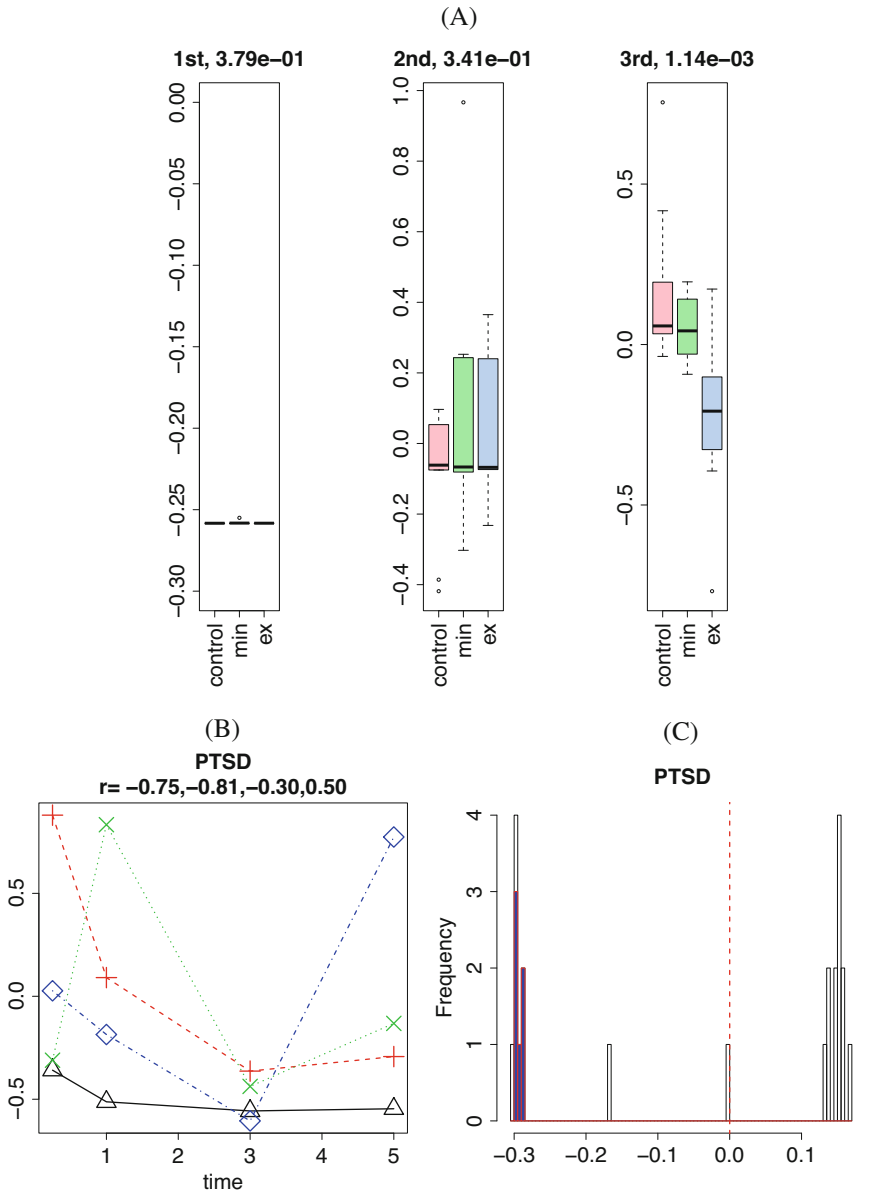

Fig. 7.32 (**a**) $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $1 \le \ell_3 \le 3$, P-values are computed by categorical regression, Eqs. (7.72) and (7.73). (**b**) $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$, open triangle: $\ell_2 = 1$, red plus symbol: $\ell_2 = 2$, green cross symbol: $\ell_2 = 3$, blue diamond: $\ell_2 = 4$. r: correlation coefficient. (**c**) Histogram of $\boldsymbol{u}_2^{(j_1)}$. Blue parts are selected ones. Vertical red broken line is 0

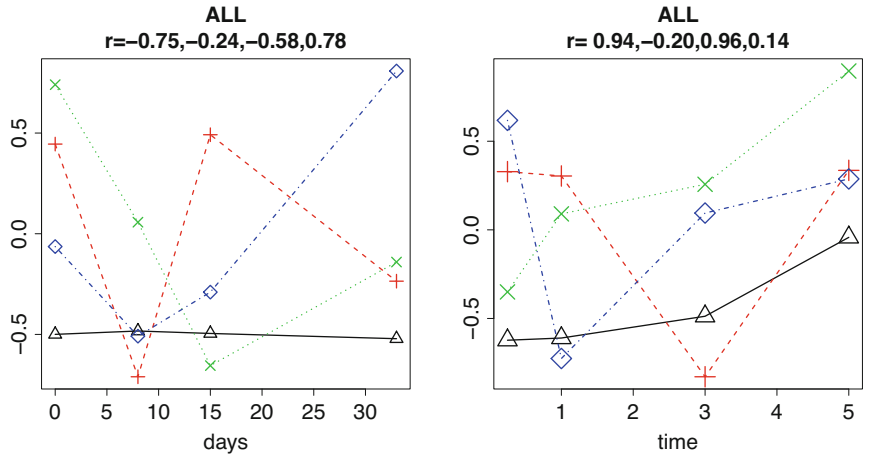

Fig. 7.33 (**a**) $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $1 \le \ell_3 \le 4$, (**b**) $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$, open triangle: $\ell_2 = 1$, red plus symbol: $\ell_2 = 2$, green cross symbol: $\ell_2 = 3$, blue diamond: $\ell_2 = 4$. r: correlation coefficient

We compute correlation coefficients between $\boldsymbol{u}_{\ell_3}^{(j_3)}$ and days after a remission induction therapy, we decide to select $\ell_3 = 4$ because it has the largest absolute value of correlation coefficient (Fig. 7.33a).

Next we try to identify $\boldsymbol{u}_{\ell_2}^{(j_2)}$ associated with significant time dependence. Figure 7.33b shows the $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$. The correlation coefficients between $\boldsymbol{u}_{\ell_2}^{(j_2)}$ and (1/4,1,3,5) are 0.94, -0.20, 0.96, and 0.14. Then $\ell_2 = 3$ with largest absolute value is selected. Then we need to find $\boldsymbol{u}_{\ell_1}^{(j_1)}$ and $\boldsymbol{u}_{\ell_5}^{(i)}$ associated with larger absolute $G(\ell_1, 3, 4, \ell_4, \ell_5)$ in order to select compounds j_1 and genes i associated with time dependence and distinction between patients and healthy controls simultaneously. In order that, we list top 20 $G(\ell_1, 3, 4, \ell_4, \ell_5)$ (Table 7.34). For ℓ_1 and ℓ_5, we decide to select those associated with top 10 Gs. As a result, $\ell_1 = 2, 3, 5, 6, 9, 10$ and $\ell_5 = 1, 2, 3, 5$ are selected. P-values are attributed to j_1 and i as

$$P_{j_1} = P_{\chi^2}\left[> \sum_{\ell_1=2,3,5,6,9,10} \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}} \right)^2 \right], \tag{7.76}$$

$$P_i = P_{\chi^2}\left[> \sum_{\ell_5=1,2,3,5} \left(\frac{u_{\ell_5 i}}{\sigma_{\ell_5}} \right)^2 \right]. \tag{7.77}$$

P-values are corrected by BH criterion and two compounds and 24 genes associated with adjusted P-values less than 0.01 are selected.

Table 7.34 Top 20 $G(\ell_1, 3, 4, \ell_4, \ell_5)$

Rank	1	2	3	4	5	6	7	8	9	10
ℓ_1	3	5	2	3	10	9	6	3	2	9
ℓ_4	4	4	4	7	4	4	4	5	4	4
ℓ_5	1	1	1	5	3	3	1	5	2	2
$G(\ell_1, 3, 4, \ell_4, \ell_5)$	260.6	−40.2	40.6	−20.9	20.7	20.4	−19.9	−18.0	16.8	−15.0

Rank	11	12	13	14	15	16	17	18	19	20
ℓ_1	8	6	14	3	3	13	12	2	1	3
ℓ_4	4	4	4	8	2	4	4	4	4	2
ℓ_5	6	4	2	5	5	4	2	3	4	1
$G(\ell_1, 3, 4, \ell_4, \ell_5)$	−13.9	13.3	13.2	−12.8	12.3	11.6	11.4	11.3	10.5	−10.5

Diabetes

Drug treated rat kidney gene expression profiles are downloaded from GEO with GEO ID GSE59913. Human diabetic kidney gene expression profile are downloaded from GEO with GEO ID GSE30122. Files used are GSE59913-GPL5425_series_matrix.txt.gz, GSE59913-GPL5426_series_matrix.txt.gz, and GSE30122_series_matrix.txt.gz. Case II tensor, $x_{ij_1j_2j_3}$, is generated as

$$x_{ij_1j_2j_3} = x_{ij_1j_2}x_{ij_3}. \tag{7.78}$$

HOSVD algorithm, Fig. 3.8, is applied to $x_{ij_1j_2j_3}$.

At first, we try to find $\boldsymbol{u}_{\ell_3}^{(j_3)}$ associated with significant distinction between four classes, normal human glomeruli, normal human kidney, normal human tubuli, and diabetic human kidney, by applying categorical regression

$$u_{\ell_3 j_3}^{(j_3)} = a_{\ell_3} + \sum_{s=1}^{4} b_{\ell_3 s}\delta_{j_3 s} \tag{7.79}$$

P-values computed by categorical regression are corrected by BH criterion. Then we found that $\ell_3 = 1, 4$ are associated with adjusted P-values less than 0.01, raw P-values of which are 2.69×10^{-9} and 1.66×10^{-9} and are selected. Figure 7.34a shows the $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $1 \le \ell_3 \le 4$.

Next we try to identify $\boldsymbol{u}_{\ell_2}^{(j_2)}$ associated with significant time dependence. Figure 7.34b shows the $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$. The correlation coefficients between $\boldsymbol{u}_{\ell_2}^{(j_2)}$ and (1/4,1,3,5) are $-0.60, -0.85, 0.53$, and 0.20. Then $\ell_2 = 2$ with largest absolute value is selected. Then we need to find $\boldsymbol{u}_{\ell_1}^{(j_1)}$ and $\boldsymbol{u}_{\ell_4}^{(i)}$ associated with larger absolute $G(\ell_1, 2, 1, \ell_4)$ or $G(\ell_1, 2, 4, \ell_4)$ in order to select compounds j_1 and genes i associated with time dependence and distinction between patients and

Table 7.35 Top 20 $G(\ell_1, 2, 1, \ell_4)$ or $G(\ell_1, 2, 4, \ell_4)$

Rank	1	2	3	4	5	6	7	8	9	10
ℓ_1	2	2	3	4	4	3	9	11	2	4
ℓ_3	1	4	1	1	4	4	1	1	1	1
ℓ_4	1	4	1	1	4	4	48	59	42	42
$G(\ell_1, 2, \ell_3, \ell_4)$	−1410	955	−75	74	−53	51	38	34	−34	34

Rank	11	12	13	14	15	16	17	18	19	20
ℓ_1	2	9	2	9	11	9	6	11	9	4
ℓ_3	4	1	1	1	1	4	1	1	4	1
ℓ_4	40	29	31	39	32	33	26	11	18	31
$G(\ell_1, 2, \ell_3, \ell_4)$	−33	33	−32	31	31	−30	−30	−29	−29	28

healthy controls simultaneously. In order that, we list top 20 $G(\ell_1, 2, 1, \ell_4)$ or $G(\ell_1, 2, 4, \ell_4)$ (Table 7.35). Because top two Gs are outstandingly large, we select $\ell_1 = 2$ and $\ell_4 = 1, 4$ associated with top two Gs.

Figure 7.34c shows $\boldsymbol{u}_2^{(j_1)}$. Fourteen outlier drugs, $u_{2j_1}^{(j_1)} > 0.13$, blue parts, are selected, by visual inspection, because P-values computed from $\boldsymbol{u}_2^{(j_1)}$ and corrected by BH criterion cannot be less than 0.01. On the other hand, P-values are attributed to ith gene as

$$P_i = P_{\chi^2}\left[> \sum_{\ell_4=1,4} \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}} \right)^2 \right] \tag{7.80}$$

P-values are corrected by BH criterion and 65 genes associated with adjusted P-values less than 0.01 are selected.

Renal Carcinoma

Drug treated rat kidney gene expression profiles are downloaded from GEO with GEO ID GSE59913. Human renal cancer gene expression profile are downloaded from GEO with GEO ID GSE40435. Files used are GSE59913-GPL5425_series_matrix.txt.gz, GSE59913-GPL5426_series_matrix.txt.gz, and GSE40435_series_matrix.txt.gz. Case II tensor, $x_{ij_1j_2j_3}$, is generated as

$$x_{ij_1j_2j_3} = x_{ij_1j_2} x_{ij_3}. \tag{7.81}$$

HOSVD algorithm, Fig. 3.8, is applied to $x_{ij_1j_2j_3}$.

At first, we try to find $\boldsymbol{u}_{\ell_3}^{(j_3)}$ associated with significant distinction between two classes, normal and cancer kidney, by applying categorical regression

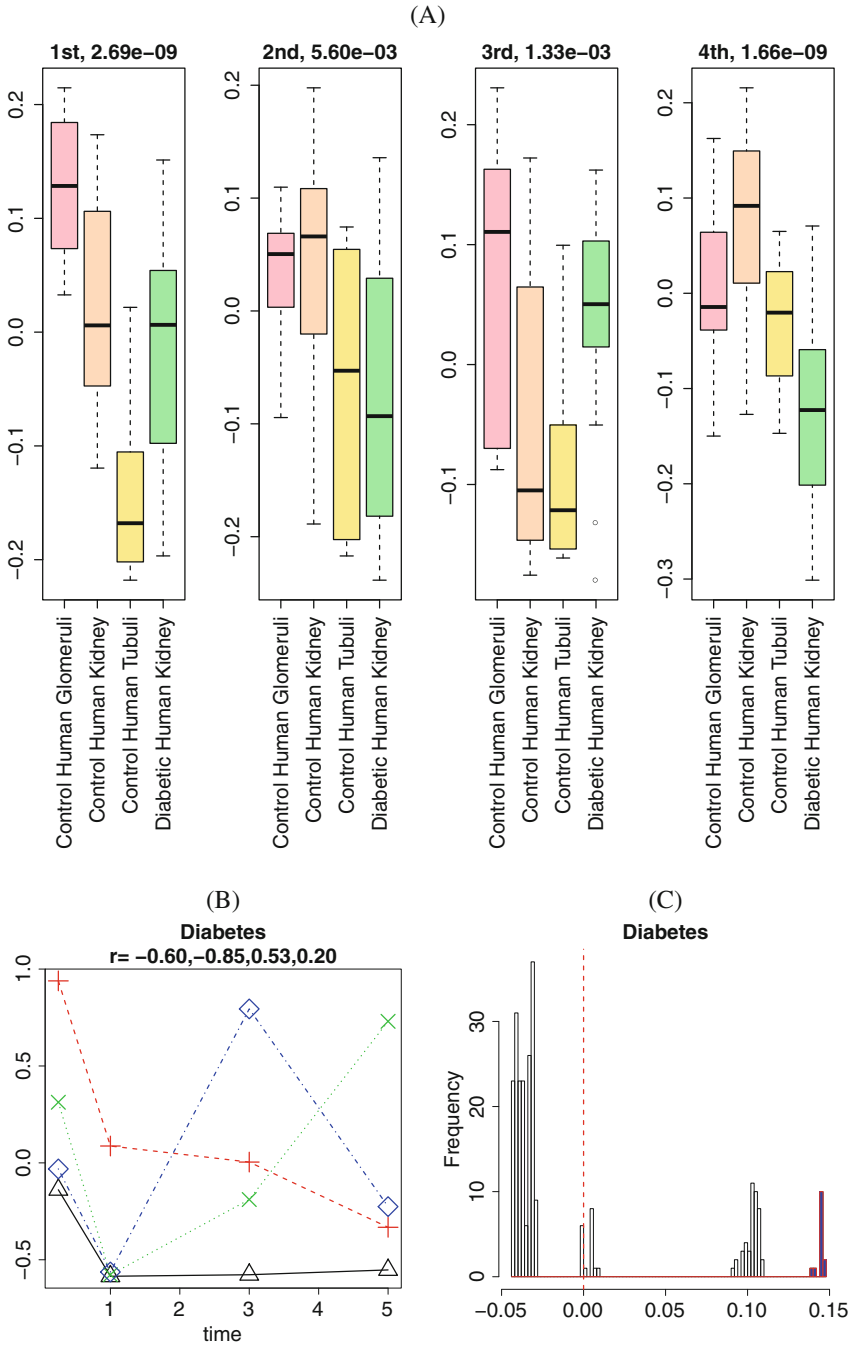

Fig. 7.34 (**a**) $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $1 \le \ell_3 \le 4$, P-values are computed by categorical regression, Eq. (7.79). (**b**) $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$, open triangle: $\ell_2 = 1$, red plus symbol: $\ell_2 = 2$, green cross symbol: $\ell_2 = 3$, blue diamond: $\ell_2 = 4$. r: correlation coefficient. (**c**) Histogram of $\boldsymbol{u}_2^{(j_1)}$. Blue parts are selected ones. Vertical red broken line is 0

$$u^{(j_3)}_{\ell_3 j_3} = a_{\ell_3} + \sum_{s=1}^{2} b_{\ell_3 s}\delta_{j_3 s} \tag{7.82}$$

P-values computed by categorical regression are corrected by BH criterion. Then we found that $\ell_3 = 13, 15, 30, 33, 35$ are associated with adjusted P-values less than 0.05, raw P-values of which are 3.4×10^{-4}, 1.1×10^{-3}, 2.7×10^{-4}, 1.1×10^{-4}, and 2.4×10^{-4} and are selected. Figure 7.35a shows the $\boldsymbol{u}^{(j_3)}_{\ell_3}$, $\ell_3 = 13, 15, 30, 33, 35$.

Next we try to identify $\boldsymbol{u}^{(j_2)}_{\ell_2}$ associated with significant time dependence. Figure 7.35b shows the $\boldsymbol{u}^{(j_2)}_{\ell_2}$, $1 \le \ell_2 \le 4$. The correlation coefficients between $\boldsymbol{u}^{(j_2)}_{\ell_2}$ and (1/4,1,3,5) are $-0.60, -0.84, 0.54$, and 0.21. Then $\ell_2 = 2$ with largest absolute value is selected. Then we need to find $\boldsymbol{u}^{(j_1)}_{\ell_1}$ and $\boldsymbol{u}^{(i)}_{\ell_4}$ associated with larger absolute $G(\ell_1, 2, \ell_3, \ell_4)$, $\ell_3 = 13, 15, 30, 33, 35$ in order to select compounds j_1 and genes i associated with time dependence and distinction between patients and healthy controls simultaneously. In order that, we list top 20 $G(\ell_1, 2, \ell_3, \ell_4)$, $\ell_3 = 13, 15, 30, 33, 35$ (Table 7.36). For top 20 Gs, it is always that $\ell_1 = 2$. On the other hand, because G gradually changes, we cannot decide threshold values. Thus, we tentatively decide that $\ell_4 = 186, 215, 233, 244, 251, 269, 274, 309, 312, 318$ associated with top 10 Gs.

Figure 7.35c shows $\boldsymbol{u}^{(j_1)}_2$. Fourteen outlier drugs, $u^{(j_1)}_{2j_1} > 0.13$, blue parts, are selected, by visual inspection, because P-values computed from $\boldsymbol{u}^{(j_1)}_2$ and corrected by BH criterion cannot be less than 0.01. On the other hand, P-values are attributed to ith gene as

$$P_i = P_{\chi^2}\left[> \sum_{\ell_4=186,215,233,244,251,269,274,309,312,318} \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}}\right)^2 \right] \tag{7.83}$$

Table 7.36 Top 20 $G(\ell_1, 2, \ell_3, \ell_4)$, $\ell_3 = 13, 15, 30, 33, 35$

Rank	1	2	3	4	5	6	7	8	9	10
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	13	13	13	13	15	15	13	13	13	15
ℓ_4	215	269	233	186	309	312	251	244	274	318
$G(\ell_1, 2, \ell_3, \ell_4)$	5.63	−5.30	5.08	−5.06	−4.84	4.78	4.66	4.61	4.57	−4.56
Rank	11	12	13	14	15	16	17	18	19	20
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	13	15	15	15	15	13	15	13	13	15
ℓ_4	289	399	336	206	363	255	375	219	342	297
$G(\ell_1, 2, \ell_3, \ell_4)$	−4.53	4.43	4.37	4.24	−4.19	−4.05	4.04	−3.97	−3.88	3.86

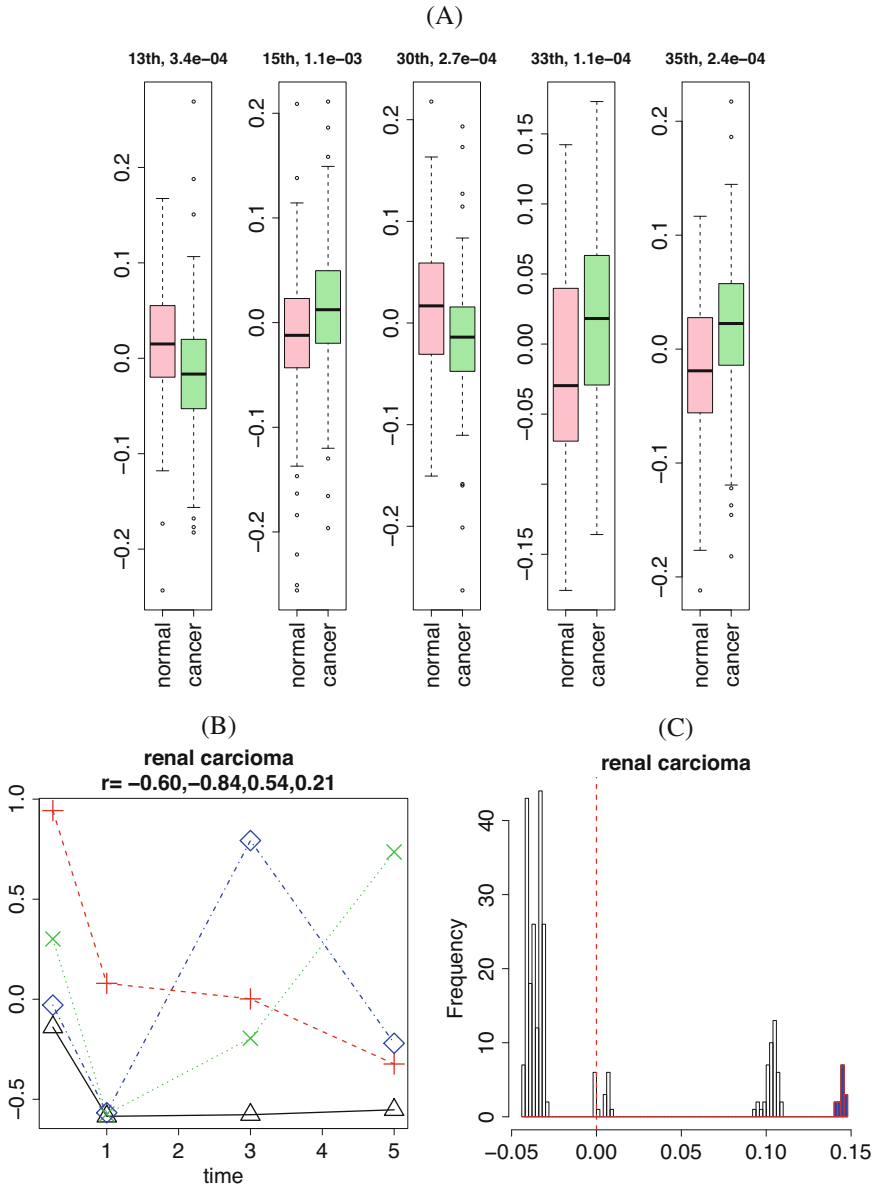

Fig. 7.35 (**a**) $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $\ell_3 = 13, 15, 30, 33, 35$, P-values are computed by categorical regression, Eq. (7.85). (**b**) $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$, open triangle: $\ell_2 = 1$, red plus symbol: $\ell_2 = 2$, green cross symbol: $\ell_2 = 3$, blue diamond: $\ell_2 = 4$. r: correlation coefficient. (**c**) Histogram of $\boldsymbol{u}_2^{(j_1)}$. Blue parts are selected ones. Vertical red broken line is 0

P-values are corrected by BH criterion and 225 genes associated with adjusted P-values less than 0.01 are selected.

Cirrhosis

Drug treated rat liver gene expression profiles are downloaded from GEO with GEO ID GSE59923. Cirrhosis patient human liver gene expression profile is downloaded from GEO with GEO ID GSE15654. File used are GSE15654_series_matrix.txt.gz, GSE59923-GPL5424_series_matrix.txt.gz, GSE59923-GPL5425_series_matrix.txt.gz, and GSE59923-GPL5426_series_matrix.txt.gz. Case II tensor, $x_{ij_1j_2j_3}$, is generated as

$$x_{ij_1j_2j_3} = x_{ij_1j_2}x_{ij_3}. \tag{7.84}$$

HOSVD algorithm, Fig. 3.8, is applied to $x_{ij_1j_2j_3}$.

At first, we try to find $\boldsymbol{u}_{\ell_3}^{(j_3)}$ associated with significant distinction between three classes, good, intermediate, and poor prognosis, by applying categorical regression

$$u_{\ell_3 j_3}^{(j_3)} = a_{\ell_3} + \sum_{s=1}^{3} b_{\ell_3 s}\delta_{j_3 s} \tag{7.85}$$

P-values computed by categorical regression are corrected by BH criterion. Then we found that $\ell_3 = 2, 6$ are associated with adjusted P-values less than 0.01, raw P-values of which are 2.3×10^{-14} and 1.0×10^{-9} and are selected. Figure 7.36a shows the $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $\ell_3 = 2, 6$.

Next we try to identify $\boldsymbol{u}_{\ell_2}^{(j_2)}$ associated with significant time dependence. Figure 7.36b shows the $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \le \ell_2 \le 4$. The correlation coefficients between $\boldsymbol{u}_{\ell_2}^{(j_2)}$ and (1/4,1,3,5) are $-0.56, -0.78, 0.52$ and 0.36. Then $\ell_2 = 2$ with largest absolute value is selected. Then we need to find $\boldsymbol{u}_{\ell_1}^{(j_1)}$ and $\boldsymbol{u}_{\ell_4}^{(i)}$ associated with larger absolute $G(\ell_1, 2, \ell_3, \ell_4)$, $\ell_3 = 2, 6$ in order to select compounds j_1 and genes i associated with time dependence and distinction between patients and healthy controls simultaneously. In order that, we list top 20 $G(\ell_1, 2, \ell_3, \ell_4)$, $\ell_3 = 2, 6$ (Table 7.37). For top 20 Gs, it is always that $\ell_1 = 2$. On the other hand, because G gradually changes, we cannot decide threshold values. Thus, we tentatively decide to select $2 \le \ell_4 \le 10$ associated with top 10 Gs.

Figure 7.36c shows $\boldsymbol{u}_2^{(j_1)}$. Twenty seven outlier drugs, $\left|u_{2j_1}^{(j_1)}\right| > 0.075$, blue parts, are selected, by visual inspection, because P-values computed from $\boldsymbol{u}_2^{(j_1)}$ and corrected by BH criterion cannot be less than 0.01. On the other hand, P-values are attributed to ith gene as

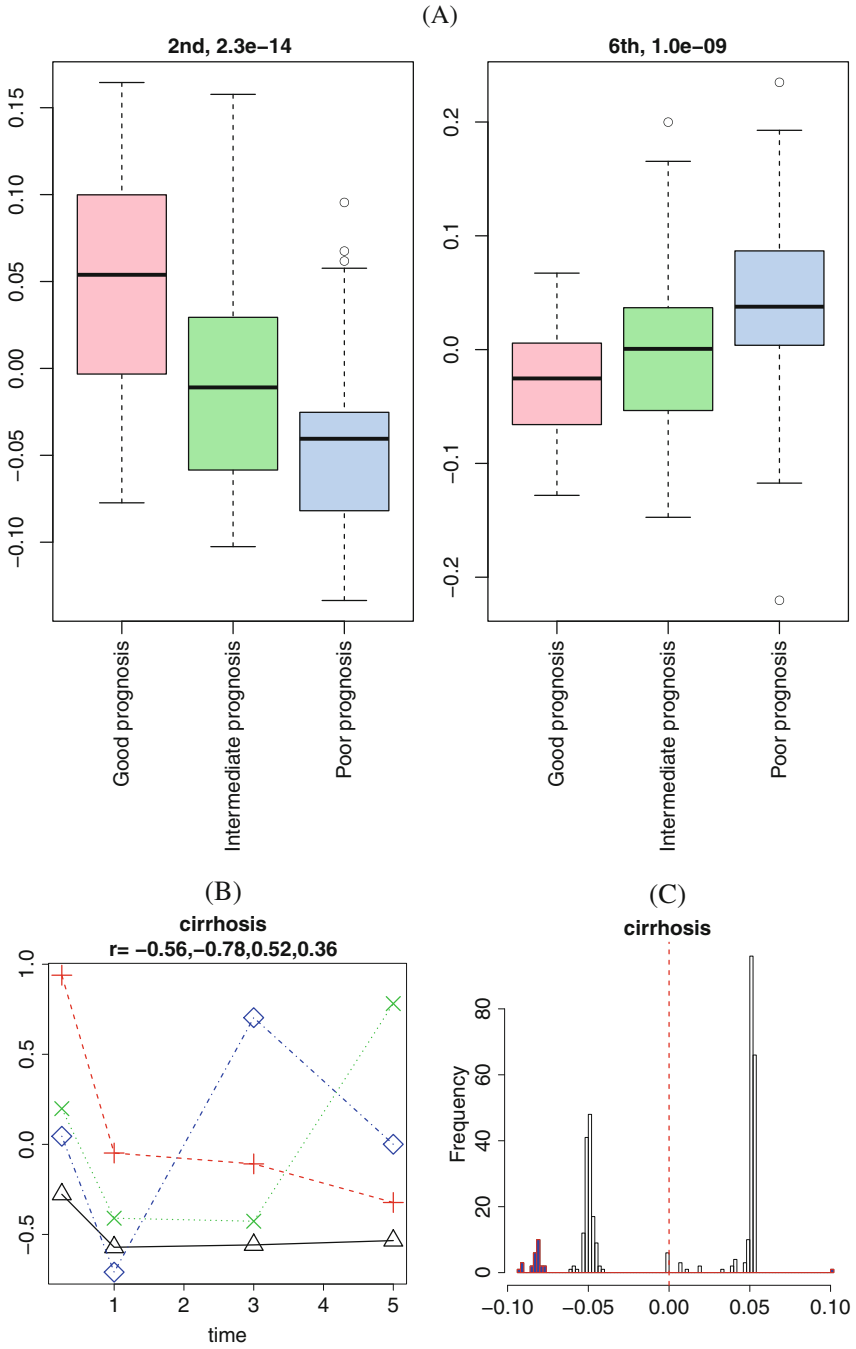

Fig. 7.36 (**a**) $\boldsymbol{u}_{\ell_3}^{(j_3)}$, $\ell_3 = 2, 6$, P-values are computed by categorical regression, Eq. (7.85). (**b**) $\boldsymbol{u}_{\ell_2}^{(j_2)}$, $1 \leq \ell_2 \leq 4$, open triangle: $\ell_2 = 1$, red plus symbol: $\ell_2 = 2$, green cross symbol: $\ell_2 = 3$, blue diamond: $\ell_2 = 4$. r: correlation coefficient. (**c**) Histogram of $\boldsymbol{u}_2^{(j_1)}$. Blue parts are selected ones. Vertical red broken line is 0

Table 7.37 Top 20 $G(\ell_1, 2, \ell_3, \ell_4)$, $\ell_3 = 2, 6$

Rank	1	2	3	4	5	6	7	8	9	10
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	2	6	6	6	6	6	2	6	2	2
ℓ_4	2	8	7	6	9	10	6	5	4	3
$G(\ell_1, 2, \ell_3, \ell_4)$	−945	310	278	194	−123	93	77	−76	−73	−67

Rank	11	12	13	14	15	16	17	18	19	20
ℓ_1	2	2	2	2	2	2	2	2	2	2
ℓ_3	6	6	6	6	6	6	6	2	6	2
ℓ_4	4	11	12	17	13	3	16	7	23	5
$G(\ell_1, 2, \ell_3, \ell_4)$	−59	49	43	40	33	−32	−31	27	25	−23

$$P_i = P_{\chi^2}\left[> \sum_{2 \le \ell_4 \le 10} \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}} \right)^2 \right] \tag{7.86}$$

P-values are corrected by BH criterion and 132 genes associated with adjusted P-values less than 0.01 are selected.

References

1. Acharya, C., Coop, A., Polli, J.E., MacKerell, A.D.: Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr. Comput. Aided Drug Des. **7**(1), 10–22 (2011). https://doi.org/10.2174/157340911793743547
2. Albrecht, M., Stichel, D., Müller, B., Merkle, R., Sticht, C., Gretz, N., Klingmüller, U., Breuhahn, K., Matthäus, F.: TTCA: an R package for the identification of differentially expressed genes in time course microarray data. BMC Bioinf. **18**(1), 33 (2017). https://doi.org/10.1186/s12859-016-1440-8
3. Anderson, A.C.: The process of structure-based drug design. Chem. Biol. **10**(9), 787–797 (2003). https://doi.org/10.1016/j.chembiol.2003.09.002. http://www.sciencedirect.com/science/article/pii/S1074552103001947
4. Bandola-Simon, J., Roche, P.A.: Dysfunction of antigen processing and presentation by dendritic cells in cancer. Mol. Immunol. (2018). http://www.sciencedirect.com/science/article/pii/S0161589018301044
5. Evans, W.E., Guy, R.K.: Gene expression as a drug discovery tool. Nat. Genet. **36**(3), 214–215 (2004). https://doi.org/10.1038/ng0304-214
6. Farhadi, T.: Advances in protein tertiary structure prediction. Biomed. Biotechnol. Res. J. (BBRJ) **2**(1), 20 (2018). https://doi.org/10.4103/bbrj.bbrj_94_17
7. Farazi, T.A., Horlings, H.M., ten Hoeve, J.J., Mihailovic, A., Halfwerk, H., Morozov, P., Brown, M., Hafner, M., Reyal, F., van Kouwenhove, M., Kreike, B., Sie, D., Hovestadt, V., Wessels, L.F., van de Vijver, M.J., Tuschl, T.: MicroRNA sequence and expression analysis in breast tumors by deep sequencing. Cancer Res. **71**(13), 4443–4453 (2011). http://cancerres.aacrjournals.org/content/71/13/4443
8. Jareborg, N., Birney, E., Durbin, R.: Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. **9**(9), 815–824 (1999). http://genome.cshlp.org/content/9/9/815.abstract

9. Jin, H.Y., Gonzalez-Martin, A., Miletic, A.V., Lai, M., Knight, S., Sabouri-Ghomi, M., Head, S.R., Macauley, M.S., Rickert, R.C., Xiao, C.: Transfection of microRNA mimics should be used with caution. Front. Genet. **6**, 340 (2015). https://www.frontiersin.org/article/10.3389/fgene.2015.00340
10. Jonic, S., Vénien-Bryan, C.: Protein structure determination by electron cryo-microscopy. Curr. Opin. Pharmacol. **9**(5), 636–642 (2009). https://doi.org/10.1016/j.coph.2009.04.006
11. Lachmann, A., Rouillard, A.D., Monteiro, C.D., Gundersen, G.W., Jagodnik, K.M., Jones, M.R., Kuleshov, M.V., McDermott, M.G., Fernandez, N.F., Duan, Q., Jenkins, S.L., Koplev, S., Wang, Z., Ma'ayan, A.: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. **44**(W1), W90–W97 (2016).https://dx.doi.org/10.1093/nar/gkw377
12. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. **39**(suppl_1), D52–D57 (2011). http://dx.doi.org/10.1093/nar/gkq1237
13. Merritt, M.A., Cramer, D.W.: Molecular pathogenesis of endometrial and ovarian cancer. Cancer Biomark. **9**(1–6), 287–305 (2011). https://doi.org/10.3233/cbm-2011-0167
14. Moustafa, A.A., Gilbertson, M.W., Orr, S.P., Herzallah, M.M., Servatius, R.J., Myers, C.E.: A model of amygdala-hippocampal-prefrontal interaction in fear conditioning and extinction in animals. Brain Cogn. **81**(1), 29–43 (2013). http://www.sciencedirect.com/science/article/pii/S0278262612001418
15. National Toxicology Program: DrugMatrix (2010). https://ntp.niehs.nih.gov/drugmatrix/index.html
16. Patalano, S., Vlasova, A., Wyatt, C., Ewels, P., Camara, F., Ferreira, P.G., Asher, C.L., Jurkowski, T.P., Segonds-Pichon, A., Bachman, M., González-Navarrete, I., Minoche, A.E., Krueger, F., Lowy, E., Marcet-Houben, M., Rodriguez-Ales, J.L., Nascimento, F.S., Balasubramanian, S., Gabaldon, T., Tarver, J.E., Andrews, S., Himmelbauer, H., Hughes, W.O.H., Guigó, R., Reik, W., Sumner, S.: Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. Proc. Natl. Acad. Sci. **112**(45), 13970–13975 (2015). https://www.pnas.org/content/112/45/13970
17. Pepper, S.D., Saunders, E.K., Edwards, L.E., Wilson, C.L., Miller, C.J.: The utility of mas5 expression summary and detection call algorithms. BMC Bioinf. **8**(1), 273 (2007). https://doi.org/10.1186/1471-2105-8-273
18. Qu, Y., He, F., Chen, Y.: Different effects of the probe summarization algorithms PLIER and RMA on high-level analysis of affymetrix exon arrays. BMC Bioinf. **11**(1), 211 (2010). https://doi.org/10.1186/1471-2105-11-211
19. Roider, H.G., Pavlova, N., Kirov, I., Slavov, S., Slavov, T., Uzunov, Z., Weiss, B.: Drug2gene: an exhaustive resource to explore effectively the drug-target relation network. BMC Bioinfor. **15**(1), 68 (2014). https://doi.org/10.1186/1471-2105-15-68
20. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., Lahr, D.L., Hirschman, J.E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I.C., Lam, D., Liberzon, A., Toder, C., Bagul, M., Orzechowski, M., Enache, O.M., Piccioni, F., Johnson, S.A., Lyons, N.J., Berger, A.H., Shamji, A.F., Brooks, A.N., Vrcic, A., Flynn, C., Rosains, J., Takeda, D.Y., Hu, R., Davison, D., Lamb, J., Ardlie, K., Hogstrom, L., Greenside, P., Gray, N.S., Clemons, P.A., Silver, S., Wu, X., Zhao, W.N., Read-Button, W., Wu, X., Haggarty, S.J., Ronco, L.V., Boehm, J.S., Schreiber, S.L., Doench, J.G., Bittker, J.A., Root, D.E., Wong, B., Golub, T.R.: A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell **171**(6), 1437–1452.e17 (2017). https://doi.org/10.1016/j.cell.2017.10.049. http://www.sciencedirect.com/science/article/pii/S0092867417313090
21. Suzuki, A., Kawano, S., Mitsuyama, T., Suyama, M., Kanai, Y., Shirahige, K., Sasaki, H., Tokunaga, K., Tsuchihara, K., Sugano, S., Nakai, K., Suzuki, Y.: DBTSS/DBKERO for integrated analysis of transcriptional regulation. Nucleic Acids Res. **46**(D1), D229–D238 (2018). http://dx.doi.org/10.1093/nar/gkx1001

22. Taguchi, Y.H.: 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), pp. 131–138 (2017). https://doi.org/10.1109/BIBE.2017.00-66
23. Taguchi, Y.H.: Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing. PLoS One **12**(8), 1–36 (2017). https://doi.org/10.1371/journal.pone.0183933
24. Taguchi, Y.H.: Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases. BMC Med. Genom. **10**(4), 67 (2017).https://doi.org/10.1186/s12920-017-0302-1
25. Taguchi, Y.H.: 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 Bioinfor. **19**(4), 99 (2018). https://doi.org/10.1186/s12859-018-2068-7
26. Taguchi, Y.H.: 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) (2018). http://www.mdpi.com/2073-4409/7/6/54
27. Taguchi, Y.H.: Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data. BMC Bioinfor. **19**(13), 388 (2019). https://doi.org/10.1186/s12859-018-2395-8
28. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. **19**, 68–77 (2015).http://dx.doi.org/10.5114/wo.2014.47136
29. Weiner, S.A., Toth, A.L.: Epigenetics in social insects: a new direction for understanding the evolution of castes. Genet. Res. Int. **2012**, 1–11 (2012). https://doi.org/10.1155/2012/609810
30. Xie, X., Luo, X., Xie, M., Liu, Y., Wu, T.: Risk of lung cancer in Parkinson's disease. Oncotarget **7**(47) (2016). https://doi.org/10.18632/oncotarget.12964
31. Yamanishi, Y., Kotera, M., Moriya, Y., Sawada, R., Kanehisa, M., Goto, S.: DINIES: drug-target interaction network inference engine based on supervised analysis. Nucleic Acids Res. **42**(W1), W39–W45 (2014). http://dx.doi.org/10.1093/nar/gku337
32. Yan, H., Bonasio, R., Simola, D.F., Liebig, J., Berger, S.L., Reinberg, D.: DNA methylation in social insects: How epigenetics can control behavior and longevity. Annu. Rev. Entomol. **60**(1), 435–452 (2015). https://doi.org/10.1146/annurev-ento-010814-020803. PMID: 25341091
33. Yoo, M., Shin, J., Kim, J., Ryall, K.A., Lee, K., Lee, S., Jeon, M., Kang, J., Tan, A.C.: DSigDB: drug signatures database for gene set analysis. Bioinformatics **31**(18), 3069–3071 (2015). http://dx.doi.org/10.1093/bioinformatics/btv313