picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Novel feature selection via kernel tensor decomposition for improved multi-omics data analysis

YH. Taguchi [1,*] and Turki Turki [2]

[1] Department of Physics, Chuo University, Tokyo 112-8551, Japan and
[2] King Abdulaziz University, Department of Computer Science, Jeddah, 21589, Saudi Arabia.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Feature selection of multi-omics data analysis remains challenging since omics data include 10^2–10^5 features. How to weight an individual omics dataset is unclear and greatly affects feature selection consequences. In this study, a recently proposed kernel tensor decomposition (KTD)-based unsupervised feature extraction (FE) was extended to integrate multi-omics datasets measured over common samples in a weight-free manner.
Results: KTD-based unsupervised FE was reformatted as the collection of kernelized tensors sharing common samples and was applied to synthetic, as well as real, datasets. The proposed advanced KTD-based unsupervised FE performed comparatively with the previously proposed KTD, as well as TD-based unsupervised FE, with reduced memory and central processing unit time. This advanced KTD method, specifically designed for multi-omics analysis, attributes P-values to features, which other multi-omics-oriented methods rarely do.
Availability: Sample R code is available in https://github.com/tagtag/MultiR/
Contact: tag@granular.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Feature selection with multi-omics datasets has always been difficult; numerous proposed methods adapted to multi-omics data analysis (Reel *et al.*, 2021; Subramanian *et al.*, 2020) include a limited number of methods that perform feature selection. Most methods fail to implement feature selection because multi-omics data analysis has a strong tendency to have a *large p small n* problem (Huynh *et al.*, 2020) to which feature selection is difficult. *Large p small n* indicates the situation where there is only a small number ($= n$) of samples with a large number ($= p$) of features. In the *large p small n* problem, feature selection is a difficult task. When features ought to have a small enough P-value under the null hypothesis, since the P-values must be heavily corrected because of multiple comparison corrections (i.e., P-values must become larger with this treatment and inevitably less significant), attributing significant P-values to individual features, even after the correction, is difficult. However, since the number of samples (i.e., conditions) is less than that of features (i.e., variables), labels or values attributed to samples may always be exactly predicted by any model (when the number of conditions is less than variables, either labels or values attributed to samples may be predicted, even if the variables are purely random numbers).

In multi-omics analysis, it is difficult to obtain large enough sample sizes since multiple observations, each of which corresponds to individual omics, must be performed. In this sense, the required cost and time are multiplied proportional to the number of omics. This often results in a smaller number of samples to which multi-omics measurements are performed when only limited experimental resources are available.

Principal component analysis (PCA) and tensor decomposition (TD)-based unsupervised feature extraction (FE) (Taguchi, 2020) have been proposed to be applied to feature selection in the *large p small n* problem. Thus, they are also fitted to feature selection in multi-omics analysis. Recently, the method has been extended to kernel TD (KTD)-based unsupervised FE (Taguchi and Turki, 2021) and has been applied to

© The Author 2015.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

integrated analysis of m6A and gene expression (Roy and Taguchi, 2021). These methods attribute P-values to features; the capability of attributing P-values to features is critically important since the significance of the selected features may be evaluated based upon attributed P-values, which other methods applicable to multi-omics datasets (Reel *et al.*, 2021; Subramanian *et al.*, 2020) rarely do. In spite of this, there are limitations of PCA, TD and KTD-based unsupervised FE when they are applied to multi-omics data analysis. In this paper, the KTD-based unsupervised FE was modified to be more suitable for multi-omics data analysis. Although only a small modification was implemented, it resulted in more flexibility during multi-omics data analysis.

2 System and Methods

2.1 A new advanced KTD-based unsupervised FE

Suppose we have K multi-omics datasets having N_k features formatted as tensors sharing sample indices $j_1, \cdots, j_m$as

$$x_{i_k j_1 j_2 \cdots j_m} \in \mathbb{R}^{N_k \times M_1 \times M_2 \times \cdots \times M_m} \tag{1}$$

where $k \in [1, K]$ and is kernelized as

$$\begin{aligned} x_{j_1 \cdots j_m j'_1 \cdots j'_m k} &= K^k(x_{i_k j_1 j_2 \cdots j_m}, x_{i_k j'_1 j'_2 \cdots j'_m}) && (2) \\ &\in \mathbb{R}^{M_1 \times \cdots \times M_m \times M_1 \times \cdots \times M_m \times K} && (3) \end{aligned}$$

where K^k is an arbitrary kernel applied to $x_{i_k j_1 j_2 \cdots j_m}$. Higher order singular value decomposition (HOSVD) (Taguchi, 2020) has been applied to $x_{j_1 \cdots j_m j'_1 \cdots j'_m k}$ resulting in eq. (4), where $G \in \mathbb{R}^{M_1 \times \cdots \times M_m \times M_1 \times \cdots \times M_m \times K}$ is a core tensor that represents the weight of individual terms composed of products of singular value vectors, $u_{\ell_s j_s} \in \mathbb{R}^{M_s \times M_s}$ and $u_{\ell_{2m+1} k} \in \mathbb{R}^{K \times K}$, which are also orthogonal matrices. Here one should notice that $u_{\ell_{2m+1} k}$ represents balance (weight) between multi-omics datasets, which usually must be pre-defined manually in the case of normal supervised learning.

Next, the $u_{\ell_s j_s}$s and $u_{\ell_{2m+1} k}$ that are of interest from the biological point of view, e.g., distinct between two classes, must be identified. Using these selected $u_{\ell_s j_s}$s, singular value vectors are derived assigned to i_ks as

$$u^{\boldsymbol{\ell}}_{i_k} = \sum_{j_1=1}^{M_1} \cdots \sum_{j_m=1}^{M_m} x_{i_k j_1 j_2 \cdots j_m} \left(\prod_{s=1}^{m} u_{\ell_s j_s} \right) \tag{5}$$

where $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_m)$.

Finally, P-values are attributed to i_k assuming that $u^{\boldsymbol{\ell}}_{i_k}$s obey a multivariate Gaussian distribution (null hypothesis) as

$$P_{i_k} = P_{\chi^2} \left[> \sum_{\boldsymbol{\ell}} \left(\frac{u^{\boldsymbol{\ell}}_{i_k}}{\sigma_{\boldsymbol{\ell}}} \right)^2 \right] \tag{6}$$

where $P_{\chi^2}[> x]$ is a cumulative χ^2 distribution where the argument is larger than x and $\sigma_{\boldsymbol{\ell}}$ is the standard deviation.

The obtained P_{i_k}s are corrected using the Benjamini-Hochberg (BH) criterion (Taguchi, 2020) and the i_ks associated with adjusted P-values less than the threshold (typically, 0.01) are selected.

2.2 Synthetic dataset

The synthetic dataset is provided as a form of tensor, $x_{ijk} \in \mathbb{R}^{N \times M \times K}$, as

$$x_{ijk} = \begin{cases} \epsilon_{ijk} + a_j & i \leq N_1, kN_1 < i \leq (k+1)N_1 \\ \epsilon_{ijk} & \text{otherwise} \end{cases} \tag{7}$$

where

$$a_j = 1 + \frac{(M-1)j_1}{M} \tag{8}$$

and $\epsilon_{ijk} \in (0, 1)$ is a uniform random real number. N is the number of variables, j is a variable that adds order to jth samples, and k is the kth omics data.

2.3 Multi-omics hepatitis B virus (HBV) vaccine dataset

Gene expression and methylation profiles were retrieved from Gene Expression Omnibus (GEO) using GSE155198 and GSE161020, respectively. For gene expression profiles, GSE155198_RAW.tar was available in the Supplementary file section of GEO. Individual files that included count number of mapped reads toward genes were collected and integrated as one file. For the methylation profiles, GSE161020_series_matrix.txt.gz, available in the GEO Series Matrix File(s) section, was used as is. The proteome dataset was obtained from ProteomeXchange (Deutsch *et al.*, 2016) using ID PXD020474. Two files named GR01,04,09,10,11,13,15,17,18,19.txt and GR02,03,05,06,07.txt were downloaded and used.

These datasets were composed of 75 samples measured in 15 individuals at five subsequent time points, i.e., 0, 1, 3, 7, and 14 days after HBV vaccine treatment. Gene expression was measured using RNA sequencing technology whereas methylation profiles were measured using microarray technology. The proteome was measured for whole blood cells (WBCs), as well as plasma. Thus, the multi-omics dataset is composed of four omics data measured for 15 individuals at five time points and is formatted as tensor $x_{i_k j_1 j_2} \in \mathbb{R}^{N_k \times 5 \times 15}$, i.e., the amount of values of i_k features of kth kind of feature measured at j_1th time points for j_2th individuals. They are standardized as $\sum_{i_k} x_{i_k j_1 j_2} = 0$ and $\sum_{i_k} x^2_{i_k j_1 j_2} = N_k$; $N_k = 687582$ ($k = 1$: methylation), 35829 ($k = 2$: gene expression), 1588 ($k = 3$: WBC proteome and $k = 4$: plasma proteome). Figure 1 illustrates how the HBV vaccination datasets are analyzed.

2.4 Kidney cancer multi-omics dataset

There were two sets of multi-omics kidney cancer data, each of which was composed of messenger RNA (mRNA) and micro RNA (miRNA) expression profiles. The first one was obtained from The Cancer Genome Atlas (TCGA) and included 253 kidney tumors and 71 normal kidneys. The second one was obtained from GEO (GSE1644117) and included 17 patients and 17 healthy controls. The method by which these two datasets were pre-processed is found in (Ng and Taguchi, 2020). The dataset is formatted as $x_{i_k j} \in \mathbb{R}^{N_k \times M}$, which represents i_ks expression of jth subjects ($k = 1$: mRNA and $k = 2$: miRNA).

3 Results

3.1 Synthetic data set

One hundred ensembles of x_{ijk} were generated and the performances were averaged. First, a linear kernel is generated as

$$x_{kjj'} = K^k(x_{ijk}, x_{i_k j' k}) = \sum_{i=1}^{N} x_{ijk} x_{ij'k} \in \mathbb{R}^{M \times M \times K} \tag{9}$$

Then, HOSVD was applied, resulting in

$$x_{kjj'} = \sum_{\ell_1=1}^{M} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 j} u_{\ell_2 j'} u_{\ell_3 k} \tag{10}$$

Then, it was noticed that u_{2j} always had the largest correlation with a_j and u_{1k} was always constant, regardless of k. Then, u_{2j} and u_{1k} were

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

$$x_{j_1\cdots j_m j'_1\cdots j'_m k} = \sum_{\ell_1=1}^{M_1}\cdots\sum_{\ell_m=1}^{M_m}\sum_{\ell_{m+1}=1}^{M_1}\cdots\sum_{\ell_{2m}=1}^{M_m}\sum_{\ell_{2m+1}=1}^{K} G(\ell_1,\cdots,\ell_m\ell_{m+1},\cdots,\ell_{2m}\ell_{2m+1})\left(\prod_{s=1}^{m} u_{\ell_s j_s}\right)\left(\prod_{s'=1}^{m} u_{\ell_{m+s'} j'_{s'}}\right) u_{\ell_{2m+1}k} \tag{4}$$

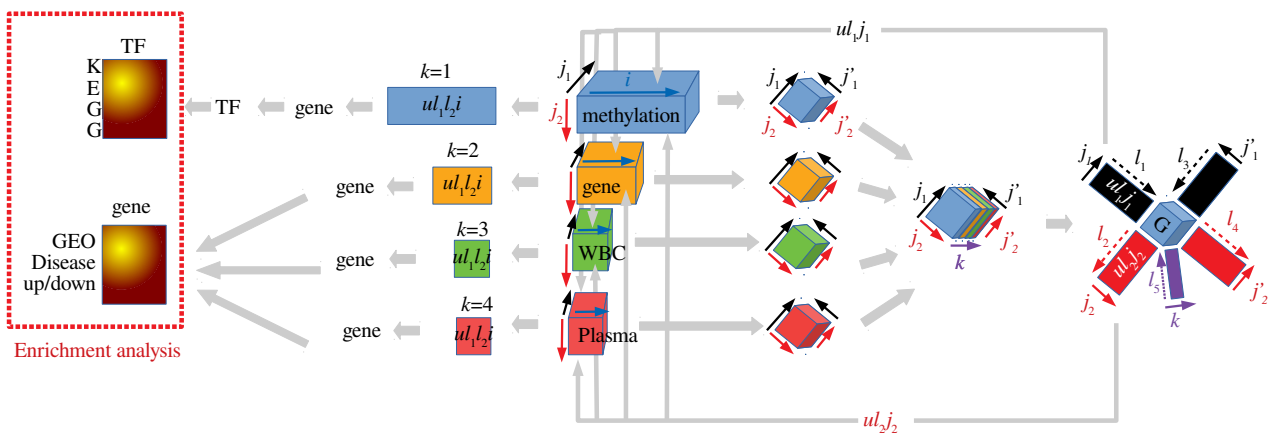

Fig. 1. Schematic figure of hepatitis B virus (HBV) vaccination data analysis. Analysis starts from the center, goes right, comes back to the center, and goes left.

employed to compute

$$u_{\ell_1\ell_2 i} = \sum_{j_1=1}^{M}\sum_{k=1}^{3} x_{ijk} u_{\ell_j} u_{\ell_2 k} \tag{11}$$

to attribute P-values

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1\ell_2 i}}{\sigma_{\ell_1\ell_2}}\right)^2\right] \tag{12}$$

P_is were corrected using the BH criterion and is associated with adjusted P_i less than 0.01 were selected. Table 1 shows the confusion matrix.

Table 1. Confusion matrix when KTD-based unsupervised FE is applied to a synthetic dataset ($N = 1000$, $N_1 = 10$, $M = 3$).

	adjusted $P_i \le 0.01$	adjusted $P_i > 0.01$
$i \le N_1$	7.06	2.94
$i > N_1$	0.04	989.96

Among 10 features associated with a_i, approximately seven features were correctly selected, whereas false positives were almost zero. Thus KTD-based unsupervised FE successfully selected features correlated with a_j.

To demonstrate how difficult this task was, linear regression was used

$$x_{ijk} = \alpha_{ik} a_j + \beta_{ik} \tag{13}$$

as an alternative method (α_{ik}, β_{ik} are regression coefficients). Since x_{ijk}s with distinct k differ from each other, distinct regression models were applied to x_{ijk} with different k. After computing BH-corrected P-values, is associated with adjusted P-values less than 0.01 were selected. Table 2 shows the confusion matrix. Essentially no features correlated with a_j were selected. Thus, regression analysis did not select features correlated with a_j.

Table 2. Confusion matrix when linear regression is applied to a synthetic dataset ($N = 1000$, $N_1 = 10$, $M = 3$).

	adjusted $P_i \le 0.01$	adjusted $P_i > 0.01$
$k = 1$		
$i \le 2N_1$	0.07	19.93
$i > 2N_1$	0.03	979.97
$k = 2$		
$i \le N_1, 2N_1 < i \le 3N_1$	0.07	19.93
Other than above	0.01	979.99
$k = 3$		
$i \le N_1, 3N_1 < i \le 4N_1$	0.09	19.91
Other than above	0.01	979.99

These results demonstrated that an apparently simple and easy problem became difficult when it is a *large p small n* problem, whereas KTD-based unsupervised FE was able to handle it to some extent. These advantages are also observed in PCA- and TD-based unsupervised FE (Taguchi, 2020).

Although the methods that did not attribute P-values to features were not of interest, since the capability of attributing P-values to features was a great advantage of KTD-based unsupervised FE, as emphasized in the Introduction, another method was employed, least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996), without attributing P-values to features, but was fitted to feature selection in a *large p small n* problem to demonstrate the difficulty of feature selection in the synthetic data. Table 3 shows the confusion matrix, which is obviously inferior to Table 1. Although the KTD-based unsupervised FE correctly selected at least 7 out of 10 features (70 %), which was correlated with a_j with essentially no false positives, lasso selected, at most, 5 out of 20 features (only 25 %), which was correlated with a_j with 2 false positives (approximately half of the true positives). This suggested that even if methods that could not attribute P-values to features were allowed to be considered, they could not outperform the KTD-based unsupervised FE. Thus, subsequently, only methods that addressed P-values to features were emphasized.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Table 3. Confusion matrix when least absolute shrinkage and selection operator (lasso) is applied to a synthetic dataset ($N = 1000, N_1 = 10, M = 3$).

	adjusted $P_i \leq 0.01$	adjusted $P_i > 0.01$
$k = 1$		
$i \leq 2N_1$	4.62	15.383
$i > 2N_1$	2.12	977.88
$k = 2$		
$i \leq N_1, 2N_1 < i \leq 3N_1$	4.70	15.30
Other than the above	2.27	977.73
$k = 3$		
$i \leq N_1, 3N_1 < i \leq 4N_1$	4.55	15.45
Other than the above	2.12	977.78

3.2 HBV vaccine

A linear kernel has been employed as

$$K^k(x_{i_k j_1 j_2}, x_{i_k j'_1 j'_2}) = \sum_{i_k=1}^{N_k} x_{i_k j_1 j_2} x_{i_k j'_1 j'_2} \tag{14}$$

Then, a tensor is added

$$x_{k j_1 j_2 j'_1 j'_2} = K^k(x_{i_k j_1 j_2}, x_{i_k j'_1 j'_2}) \in \mathbb{R}^{4\times5\times15\times5\times15}. \tag{15}$$

Applying HOSVD to $x_{k j_1 j_2 j'_1 j'_2}$ results in

$$x_{k j_1 j_2 j'_1 j'_2} = \sum_{k=1}^{4} G(\ell_1\ell_2\ell_3\ell_4\ell_5) u_{\ell_1 j_1} u_{\ell_2 j_2} u_{\ell_3 j'_1} u_{\ell_4 j'_2} u_{\ell_5 k} \tag{16}$$

where $G \in \mathbb{R}^{5\times15\times5\times15\times4}$, $u_{\ell_1 j_1}, u_{\ell_3 j'_1} \in \mathbb{R}^{5\times5}$, $u_{\ell_2 j_4}, u_{\ell_4 j'_2} \in \mathbb{R}^{15\times15}$, and $u_{\ell_5 k} \in \mathbb{R}^{4\times4}$. Since $x_{k j_1 j_2 j'_1 j'_2} = x_{k j'_1 j_2 j_1 j'_2}$ and $x_{k j_1 j_2 j'_1 j'_2} = x_{k j_1 j'_2 j'_1 j_2}$ because of symmetry, $G(\ell_1\ell_2\ell_3\ell_4\ell_5) = G(\ell_1\ell_4\ell_3\ell_2\ell_5) = G(\ell_3\ell_2\ell_1\ell_4\ell_5)$, $\{u_{\ell_1 j_1}\} = \{u_{\ell_3 j'_1}\}$, and $\{u_{\ell_2 j_2}\} = \{u_{\ell_4 j'_2}\}$.

The singular value vectors of interest were as follows.

- $u_{\ell_1 j_1}$ and $u_{\ell_3 j'_1}$ should be significantly dependent upon time points corresponding to j_1 and j'_1.
- $u_{\ell_2 j_2}$ and $u_{\ell_4 j'_2}$ should be independent of individuals corresponding to j_2 and j'_2.
- $u_{\ell_5 k}$ should be common between distinct omics measurements.

As a result, $\ell_1 = \ell_3 = 2, \ell_2 = \ell_4 = 1, \ell_5 = 1$ satisfies the required conditions (Fig. 2). Then,

$$u_{i_k} = \sum_{j_1=1}^{5} \sum_{j_2=1}^{15} x_{i_k j_1 j_2} u_{2 j_1} u_{1 j_2} \tag{17}$$

is computed and

$$P_{i_k} = P_{\chi^2}\left[> \left(\frac{u_{i_k}}{\sigma_{21}}\right)^2 \right] \tag{18}$$

The computed P_{i_k}s were corrected using the BH criterion and i_ks associated with either $P_{i_k} < 0.01$ (for gene expression and methylation) or with $P_{i_k} < 0.05$ (for the two proteomes) were successfully selected (the full list of selected features are available in Supplementary Data S1–S4).

To validate genes selected by KTD-based unsupervised FE, selected genes were uploaded to Enrcihr (Kuleshov *et al.*, 2016). At first,

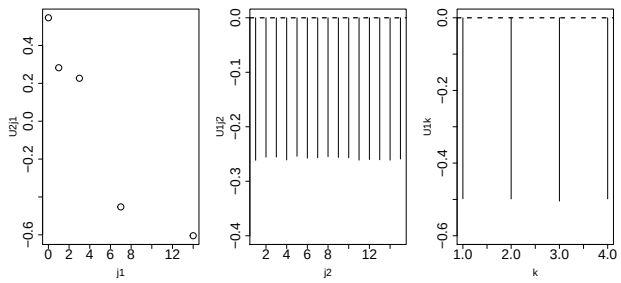

Fig. 2. Left: u_{2j_1}, middle: u_{1j_1}, right: u_{1k} when HOSVD is applied to a linear kernel computed using hepatitis B virus (HBV) vaccine data. The Pearson correlation coefficient between j_1 and u_{2j_1} is -0.94 ($P = 0.02$).

1335 genes associated with 2077 methylation probes selected by KTD-based unsupervised FE were uploaded. Many transcription factors (TFs) significantly targeted these 1335 genes (Table 4). Then, these 21 TFs were again uploaded to Enrichr; Table S1 shows the top ranked 10 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in the "KEGG 2019 HUMAN" Enrichr category (the full list is available as Supplementary Data S1). It was clear that these data included many biologically reasonable KEGG pathways (for details, see Discussion section below).

Table 4. Transcription factors (TFs) enriched in "ChEA 2016" Enrichr category (adjusted P-values < 0.05) when 1335 genes associated with 2077 methylation probes selected by KTD-based unsupervised FE are considered (the full list is available in Supplementary Data S1).

TFs	ZNF217, TCF4, STAT3, SMARCD1, WT1, FOXA2, PAX3-FKHR, SMAD4, SMAD3, SOX9, TFAP2C, YAP1, AR, SOX2, CTNNB1, VDR, PIAS1, TEAD4, MITF, HNF4A, SUZ12

Next, 8 genes associated with 11 probes identified as differentially expressed genes (DEGs) when gene expression profiles were considered (Table 5), were uploaded to Enrichr.

Table S2 shows the top-ranked 10 GEO profiles enriched in the Enrichr "Disease Perturbations from GEO up/down" category (the full list is available as Supplementary Data S2).

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Table 5. Eight genes associated with 11 probes identified as differentially expressed genes (DEGs) when gene expression profiles are considered.

Gene symbols	S100A9, CD74, hba1, ACTB, HBB, HBA2, MALAT1, COX1

Although many neurodegenerative profiles not apparently related to the HBV vaccine are listed, their rationalization is included in the Discussion section.

Finally, two set of proteins identified as DEGs in WBC and plasma profiles (Table 6) were uploaded to Enrichr.

Table 6. Proteins identified as differentially expressed genes (DEGs) when gene expression profiles in the proteome are considered.

WBC	HIST1H2BJ, HIST2H2BF, HIST1H2BG, HIST1H2BB, HIST1H2BD, ACTG1, HIST1H2BL, HIST1H2BN, PFN1, HIST1H2BK, HIST3H2BB, ACTB, HBB, HBA2, HIST1H2BA, HIST1H2BI, HIST1H2BC, HIST1H2BO, HIST2H2BE, HIST1H2BM, HBA1, HIST1H2BF, HIST1H2BE, HIST1H2BH
Plasma	FGA, HP, GSN, ALB, FGG, IGLL5, APOA1, SERPINA1, ORM1, TF, GC, CP, C4A, CSF3R, A2M, HPX, HRG, A1BG, CFH, APOB, C3, CLEC14A

Tables S3 and S4 show the top-ranked 10 enriched GEO profiles identified in the Enrichr "Disease Perturbations from GEO up/down" category when proteins for the WBC and plasma sections in Table 6 are uploaded to Enrichr (the full list is available as Supplementary Data S3 and S4). Other than the enriched reasonable hepatitis-related GEO profiles, some neurodegenerative disease-related GEO profiles are also enriched in Table S3. These data are discussed in the Discussion section.

Since there are many biological enrichments listed in Tables S1, S2, S3 and S4, the genes and proteins selected in this section may not be artifacts, but have a biological basis.

3.3 Kidney cancer

A linear kernel was employed as

$$K^k(x_{i_k j}, x_{i_k j'}) = \sum_{i_k=1}^{N_k} x_{i_k j} x_{i_k j'} \in \mathbb{R}^{N_k \times M} \tag{19}$$

For the first, i.e. TCGA dataset, $M = 324$, whereas for the second, i.e. the GEO dataset, $M = 37$. Then, a tensor was added

$$x_{kjj'} = K^k(x_{i_k j}, x_{i_k j'}) \in \mathbb{R}^{2 \times M \times M}. \tag{20}$$

Applying HOSVD to $x_{kjj'}$ results in

$$x_{kjj'} = \sum_{k=1}^{2} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 j} u_{\ell_2 j'} u_{\ell_3 k} \tag{21}$$

where $G \in \mathbb{R}^{M \times M \times 2}$, $u_{\ell_1 j}, u_{\ell_2 j'} \in \mathbb{R}^{M \times M}$, and $u_{\ell_3 k} \in \mathbb{R}^{2 \times 2}$. Since $x_{kjj'} = x_{kj'j}$ because of symmetry, $G(\ell_1 \ell_2 \ell_3) = G(\ell_1 \ell_3 \ell_3)$, $\{u_{\ell_1 j}\} = \{u_{\ell_2 j'}\}$. Then, singular value vectors were identified such that $u_{\ell_1 j}$ and $u_{\ell_2 j'}$ were significantly distinct between healthy controls and patients. As a result, $\ell_1 = \ell_2 = 2$ satisfies the required conditions (Fig. 3). Then,

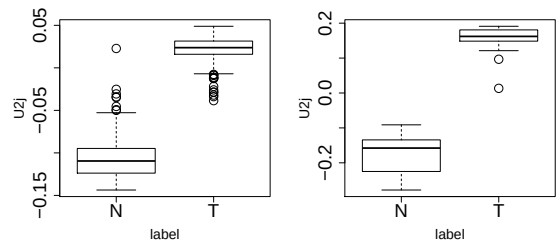

Fig. 3. Boxplot of u_{2j} when higher order singular value decomposition (HOSVD) is applied to a linear kernel computed using kidney cancer data. Left: TCGA, $P = 8.49 \times 10^{-47}$, right: GEO, $P = 4.07 \times 10^{-17}$. P-values are when t test was applied to u_{2j}. T:tumors, N:normal kidneys.

$$u_{i_k} = \sum_{j=1}^{M} x_{i_k j} u_{2j} \tag{22}$$

is computed and

$$P_{i_k} = P_{\chi^2}\left[> \left(\frac{u_{i_k}}{\sigma_2} \right)^2 \right] \tag{23}$$

The computed P_{i_k}s were corrected using the BH criterion and i_ks associated with $P_{i_k} < 0.01$ were successfully selected. hsa-mir-200c and hsa-mir-141 were selected from TCGA and hsa-miR-141, hsa-miR-210, and hsa-miR-200c were selected from GEO. Thus, these miRNAs are highly coincident with each other and even more coincident with each other than in previous works (Ng and Taguchi, 2020; Taguchi and Turki, 2021) where TD, as well as KTD-based unsupervised FE has been applied to TCGA and GEO datasets. For mRNA, there were five common genes between the TCGA and GEO datasets(Table 7, $P = 6.7 \times 10^{-5}$, Odds ratio: 13.13).

Table 7. Confusion matrix of selected mRNAs between The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets

		GEO	
		$P > 0.01$	$P < 0.01$
TCGA	$P > 0.01$	17269	101
	$P < 0.01$	65	5

$P = 6.7 \times 10^{-5}$, Odds ratio: 13.13

4 Discussion

There are several advantages in the proposed implementation of KTD-based unsupervised FE if compared to that of previously proposed versions of KTD-based unsupervised FE (Taguchi and Turki, 2021), as well as TD-based unsupervised FE (Taguchi, 2020) in the context of application to the integration of multi-omics datasets. When the original implementation of TD-based unsupervised FE (Taguchi, 2020) is applied to the integration

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Table 8. Correlation between hepatitis B virus (HBV) vaccine experiment kernels. Upper triangle: Pearson correlation, lower triangle: P-values.

—	Gene expression	WBC	Plasma
Gene expression	—	0.1279405	0.2192163
WBC	1.34×10^{-11}	—	0.4384998
Plasma	1.52×10^{-31}	8.98×10^{-131}	—

of multi-omics data, e.g. $x_{ij} \in \mathbb{R}^{N \times M}$ and $x_{hj} \in \mathbb{R}^{H \times M}$, which corresponds to ith or hth omics data of jth sample, HOSVD is applied to

$$x_{ih} = \sum_j x_{ij} x_{hj} \in \mathbb{R}^{N \times H} \tag{24}$$

Since $H, N \gg M$, this was not an effective implementation. When KTD-based unsupervised FE (Taguchi and Turki, 2021) is applied, HOSVD is applied to

$$x_{jj''} = \sum_{j'} \left(\sum_i x_{ij} x_{ij'} \right) \left(\sum_h x_{hj'} x_{hj''} \right) \in \mathbb{R}^{M \times M} \tag{25}$$

Thus, this drastically reduced the required memory and central processing unit (CPU) time. Nevertheless, it was unclear how more than two omics datasets could be integrated. In the implementation introduced in this paper,

$$x_{jj'k} = \sum_i x_{ijk} x_{ij'k} \in \mathbb{R}^{M \times M \times K} \tag{26}$$

which corresponds to the ith measurement of the kth omics data of the jth sample, where K is total number of omics datasets. Since HOSVD is applied to $x_{jj'k}$, any number of multi-omics datasets may be handled. This slight modification drastically increased the ability of KTD-based unsupervised FE that handle multi-omics datasets.

Kernels K^k are highly correlated between mRNA ($k = 1$) and miRNA ($k = 2$) for kidney cancer data (Li *et al.*, 2018) (Fig.4). Kernels K^k for gene expression profiles and the proteome were also highly correlated when HBV vaccine experiments were considered (Table 8). Thus, it was obvious that the current formalism was very effective in identifying the coincidence between individual omics (in this case, mRNA, miRNA, and the proteome).

Although Table S1 did not seem to be related to the HBV vaccine directly, there were indirect reasonable relationships. The Hippo signaling pathway has recently been identified to be related to the immune system (Hong *et al.*, 2018). "Hepatitis B" is also significantly enriched (adjusted P-value is 7.79×10^{-3}, but is not within the top-ranked 10 KEGG pathways in Table S1). Since patients with diabetes have a higher risk of HBV infection (Khalili *et al.*, 2015), it is reasonable that the two KEGG pathways, "AGE-RAGE signaling pathway in diabetic complications" and "Maturity onset of diabetes in the young" are enriched. Although there are multiple cancer types enriched in Table S1, many cancer types other than liver cancers are related to risk of HBV infection (Song *et al.*, 2019).

Although genes identified as DEGs in HBV vaccination are also enriched in various neurodegenerative diseases other than hepatitis (e.g. in Tables S2 and S3), this is reasonable because viral hepatitis has once been reported to be related to Parkinson's disease (Pakpoor *et al.*, 2017). There is also known associations between hepatic functions and plasma amyloid-β levels (Wang *et al.*, 2016); cirrhosis patients with HBV infection have higher plasma Aβ40 and Aβ42 levels than patients with HBV-negative cirrhosis. More directly, Ji *et al.* (2018) have reported that hepatitis B core

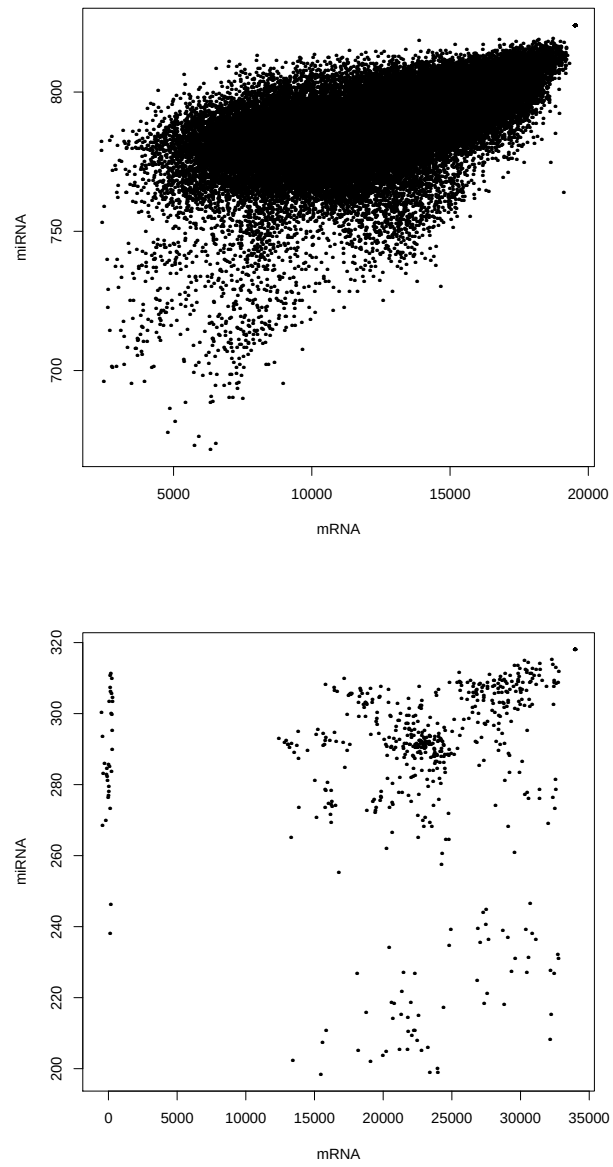

Fig. 4. Scatter plot of kernels K^k between messenger RNA (mRNA) and micro RNA (miRNA). Upper: The Cancer Genome Atlas (TCGA) (Pearson correlation coefficient= 0.627, $P = 0.00$), Lower: Gene Expression Omnibus (GEO)(Pearson correlation coefficient= 0.349, $P = 2.32 \times 10^{-34}$).

VLP-based mis-disordered tau vaccine alleviates cognitive deficits and neuropathology progression in a Tau.P301S mouse model of Alzheimer's disease. Thus, enrichment of neurodegenerative disease-related genes in identified DEGs was not an artifact, but possible evidence that KTD-based unsupervised FE detected side effects caused by vaccinations.

For comparisons with other univariate conventional tools, such as limma (Ritchie *et al.*, 2015) and sam (Tusher *et al.*, 2001), these tools cannot be used since they are for categorical classes and cannot be applied to HBV vaccination data since HBV vaccination data is only associated with time points, not categorical data. Thus, although regression analysis was attempted for the synthetic dataset, there were no features correlated with dates. Thus, there were no univariate feature selection

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

methods applicable to HBV vaccination data that could identify features correlated with date. For the kidney cancer datasets, it has been extensively demonstrated that univariate conventional tools, such as limma (Ritchie *et al.*, 2015) and sam (Tusher *et al.*, 2001), cannot compete with the TD-based unsupervised FE (Ng and Taguchi, 2020; Taguchi and Turki, 2021). Thus, no univariate feature selection method was identified that was superior to TD-based unsupervised FE when applied to kidney cancer datasets.

The performance of the proposed method has not been compared to other existing multi-omics oriented methods (Reel *et al.*, 2021; Subramanian *et al.*, 2020) because no suitable methods have been identified to be compared. First of all, most of the recently proposed cutting-edge methods adapted to multi-omics analysis are specific to high throughput sequencing (HTS) architecture; for example, MKpLMM (Li *et al.*, 2019) requires genomic coordinates, which are not available for the present datasets. csaw (Lun and Smyth, 2015) requires a bed file, which is also not available for the present datasets. Since the purpose of the present study was to propose more flexible methods that were not specific to HTS architecture and the datasets employed in this study were not obtained using HTS, then these were not suitable methods to be compared with the presently proposed implementation of KTD-based unsupervised FE. In addition to this, other methods that are not specific to HTS lack the statistical validation of feature selection (i.e., no ability to attribute P-values to features). Although DIABLO (Singh *et al.*, 2019) is not specific to HTS architecture and has feature selection ability, there is no functionality to attribute P-values to individual features; thus, features cannot be selected based upon statistical significance and, therefore, DIABLO is out of scope of this study. FSMKL (Seoane *et al.*, 2013), which is also not specific to HTS architecture, has the ability to add statistical scores to individual features; however, selecting features based upon statistical scores is not effective. In this sense, to our knowledge, there are no other multi-omics oriented feature selection methods that satisfy the following requirements

- Not specific to HTS architecture
- Attributes P-values to individual features to evaluate statistical significance
- Can handle more than or equal to three kinds of omics data
- Applicable to severe *large p small n* (typically, $p/n \sim 10^2$ or more) problems

For example, for the *large p small n* problem, although Subramanian *et al.* (2020) have summarized existing machine learning methods for multi-omics analysis, typically they are applied to studies including as many as 10^2 samples; thus, they cannot be regarded as a severe *large p small n* problem. As a result, the performance of the KTD-based unsupervised FE was not directly compared to other existing methods that satisfy all the above conditions.

HBV vaccination data were selected to demonstrate the superior power of advanced KTD-based unsupervised FE because of the difficulty of the problem. Since vaccination must be given to healthy people, side effects must be minimized (Jacobson *et al.*, 2001); actually, since vaccination is essentially an infection with a weaker pathogen, its effect is inevitably weak. As expected, a very limited number of features (genes and proteins) were selected using the advanced KTD-based unsupervised FE proposed in this article, whereas conventional linear regression analysis did not attribute significant P-values to any features. This suggests that the advanced KTD-based unsupervised FE proposed in this study has superior ability to select features when it is applied to difficult multi-omics datasets.

5 Conclusion

In this paper, an advanced KTD-based unsupervised FE was introduced, which was modified to be applied to feature selection in multi-omics data analysis that is often very difficult, that of the *large p small n* problem. It was successfully applied to synthetic, as well as two real datasets, and attributed significant P-values to features with reduced CPU time and memory, even when it was applied to integrated analysis of more than two multi-omics datasets. Although the modification from the previously proposed KTD-based unsupervised FE was not significant, this slight modification was successful when it was applied to feature selection of multi-omics data analysis, which often becomes difficult feature selection problems in a *large p small n* problem.

Funding

This work was supported by Japan Society for the Promotion of Science, KAKENHI [grant numbers 19H05270, 20K12067, 20H04848] to YHT.

References

Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D. S., Bernal-Llinares, M., Okuda, S., Kawano, S., Moritz, R. L., Carver, J. J., Wang, M., Ishihama, Y., Bandeira, N., Hermjakob, H., and Vizcaíno, J. A. (2016). The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, **45**(D1), D1100–D1106.

Hong, L., Li, X., Zhou, D., Geng, J., and Chen, L. (2018). Role of hippo signaling in regulating immunity. *Cellular & Molecular Immunology*, **15**(12), 1003–1009.

Huynh, P.-H., Nguyen, V. H., and Do, T.-N. (2020). Improvements in the large p, small n classification issue. *SN Computer Science*, **1**(4).

Jacobson, R. M., Swan, A., Adegbenro, A., Ludington, S. L., Wollan, P. C., and Poland, G. A. (2001). Making vaccines more acceptable - methods to prevent and minimize pain and other common adverse events associated with vaccines. *Vaccine*, **19**(17), 2418–2427.

Ji, M., xiu Xie, X., qun Liu, D., lin Yu, X., Zhang, Y., Zhang, L.-X., wei Wang, S., ru Huang, Y., and tian Liu, R. (2018). Hepatitis b core VLP-based mis-disordered tau vaccine elicits strong immune response and alleviates cognitive deficits and neuropathology progression in tau.p301s mouse model of alzheimer's disease and frontotemporal dementia. *Alzheimer's Research & Therapy*, **10**(1).

Khalili, M., Lombardero, M., Chung, R. T., Terrault, N. A., Ghany, M. G., Kim, W. R., Lau, D., Lisker-Melman, M., Sanyal, A., Lok, A. S., and for the HBRN (2015). Diabetes and prediabetes in patients with hepatitis b residing in north america. *Hepatology*, **62**(5), 1364–1374.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, **44**(W1), W90–W97.

Li, J., Lu, Q., and Wen, Y. (2019). Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data. *Bioinformatics*, **36**(6), 1785–1794.

Li, X., Yu, X., He, Y., Meng, Y., Liang, J., Huang, L., Du, H., Wang, X., and Liu, W. (2018). Integrated analysis of MicroRNA (miRNA) and mRNA profiles reveals reduced correlation between MicroRNA and target gene in cancer. *BioMed Research International*, **2018**, 1–15.

Lun, A. T. and Smyth, G. K. (2015). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research*, **44**(5), e45–e45.

Ng, K.-L. and Taguchi, Y.-H. (2020). Identification of miRNA signatures for kidney renal clear cell carcinoma using the tensor-decomposition method. *Scientific Reports*, **10**(1).

Pakpoor, J., Noyce, A., Goldacre, R., Selkihova, M., Mullin, S., Schrag, A., Lees, A., and Goldacre, M. (2017). Viral hepatitis and parkinson disease. *Neurology*, **88**(17), 1630–1633.

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, **49**, 107739.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

microarray studies. *Nucleic Acids Research*, **43**(7), e47.

Roy, S. S. and Taguchi, Y.-H. (2021). Identification of genes associated with altered gene expression and m6a profiles during hypoxia using tensor decomposition based unsupervised feature extraction. *Scientific Reports*, **11**(1), 8909.

Seoane, J. A., Day, I. N. M., Gaunt, T. R., and Campbell, C. (2013). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, **30**(6), 838–845.

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., and Cao, K.-A. L. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, **35**(17), 3055–3062.

Song, C., Lv, J., Liu, Y., Chen, J. G., Ge, Z., Zhu, J., Dai, J., Du, L.-B., Yu, C., Guo, Y., Bian, Z., Yang, L., Chen, Y., Chen, Z., Liu, J., Jiang, J., Zhu, L., Zhai, X., Jiang, Y., Ma, H., Jin, G., Shen, H., Li, L., Hu, Z., and for the China Kadoorie Biobank Collaborative Group (2019). Associations Between Hepatitis B Virus Infection and Risk of All Cancer Types. *JAMA Network Open*, **2**(6), e195718–e195718.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, **14**, 1177932219899051. PMID: 32076369.

Taguchi, Y.-h. (2020). *Unsupervised Feature Extraction Applied to Bioinformatics*. Springer International Publishing.

Taguchi, Y.-H. and Turki, T. (2021). Mathematical formulation and application of kernel tensor decomposition based unsupervised feature extraction. *Knowledge-Based Systems*, **217**, 106834.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**(9), 5116–5121.

Wang, Y.-R., Wang, Q.-H., Zhang, T., Liu, Y.-H., Yao, X.-Q., Zeng, F., Li, J., Zhou, F.-Y., Wang, L., Yan, J.-C., Zhou, H.-D., and Wang, Y.-J. (2016). Associations between hepatic functions and plasma amyloid-beta levels—implications for the capacity of liver in peripheral amyloid-beta clearance. *Molecular Neurobiology*, **54**(3), 2338–2344.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture