Bioinformatics, 34(24), 2018, 4151–4158
doi: 10.1093/bioinformatics/bty490
Advance Access Publication Date: 21 June 2018
Original Paper

OXFORD

Genome analysis

GIFT: Guided and Interpretable Factorization for Tensors with an application to large-scale multi-platform cancer analysis

Jungwoo Lee†, Sejoon Oh† and Lee Sael*

Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on December 31, 2017; revised on June 12, 2018; editorial decision on June 13, 2018; accepted on June 18, 2018

Abstract

Motivation: Given multi-platform genome data with prior knowledge of functional gene sets, how can we extract interpretable latent relationships between patients and genes? More specifically, how can we devise a tensor factorization method which produces an interpretable gene factor matrix based on functional gene set information while maintaining the decomposition quality and speed?
Results: We propose GIFT, a **G**uided and **I**nterpretable **F**actorization for **T**ensors. GIFT provides interpretable factor matrices by encoding prior knowledge as a regularization term in its objective function. We apply GIFT to the PanCan12 dataset (TCGA multi-platform genome data) and compare the performance with P-Tucker, our baseline method without prior knowledge constraint, and Silenced-TF, our naive interpretable method. Results show that GIFT produces interpretable factorizations with high scalability and accuracy. Furthermore, we demonstrate how results of GIFT can be used to reveal significant relations between (cancer, gene sets, genes) and validate the findings based on literature evidence.
Availability and implementation: The code and datasets used in the paper are available at https://github.com/leesael/GIFT.
Contact: saellee@snu.ac.kr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Increasing number of multi-platform genome data of a single person, e.g. a cancer patient, are being generated. These data describe different biological aspects of a person and need to be integratively analyzed to obtain a holistic view. However, due to the complexity of the problem, the results of existing methods are difficult to interpret and often do not scale to larger data (Thomas and Sael, 2015). Interpretability is important for discoveries, and scalability is also important as the size of data rapidly increase.

1.1 Integrative genomic data analysis for cancer studies

The Cancer Genome Atlas (TCGA) have reported several integrated genome-wide studies of cancer data. In 2013, TCGA published the PanCan12 dataset that includes multi-platform genomic information of 12 tumor types (Weinstein *et al.*, 2013). The dataset has boosted many genomic cancer analyses (Anaya *et al.*, 2016; Riaz *et al.*, 2017) including the original TCGA multi-platform data analysis (Hoadley *et al.*, 2014). Hoadley *et al.* (2014) utilizes cluster-of-cluster analysis (COCA) approach for stratification of the Pancan12 dataset. COCA is a two-step approach that clusters against already clustered individual data types. Although the method is applicable for large data, the two-step process makes it difficult to trace back and interpret the results. Multi-kernel methods are also multi-step approaches that first generate individual kernels from each data type, then learn multi-kernels, and finally apply the multi-kernels to cluster or classify (Thomas and Sael, 2017). Although kernel-based methods are highly accurate, interpretability is lost in the generation

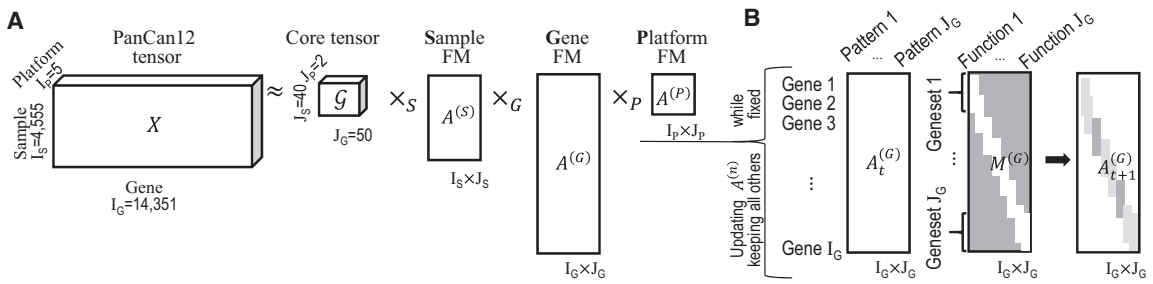

Fig. 1. An overview of a PanCan12 tensor factorization via GIFT. (**A**) shows a PanCan12 tensor factorization and (**B**) shows a factor matrix (FM) $A_t^{(G)}$ computation at time $t+1$ constrained on a mask matrix $M^{(G)}$

of the kernels. Another integrative method widely used by TCGA is PARADIGM (Kandoth *et al.*, 2013; Vaske *et al.*, 2010). The method is based on a Bayesian network inference, which depends on the biological pathway and protein expression data used. Due to these requirements, it is often applied to a small number of genes.

1.2 Matrix/tensor mining methods

Matrix factorization methods, such as the non-negative matrix factorization (NMF), are broadly used across multiple domains to analyze data represented as matrices. NMF was used extensively by the TCGA group (Hoadley *et al.*, 2014; Kandoth *et al.*, 2013; Koboldt *et al.*, 2012) and others (Hofree *et al.*, 2013; Kim *et al.*, 2015; Zhu *et al.*, 2017) for studying single-platform genome analyses such as somatic mutations or gene expressions.

Natural extensions of single data type modeled as matrices to multi-platform data are tensors, i.e. multi-dimensional arrays. Tensors are widely applied to represent many real-world data such as movie rating and network traffic data expressed as 3-order tensors with three modes (movie—user—time) and three modes (source IP—destination IP—time), respectively. Multi-platform genome data can also be represented as a 3-order tensor that contains the experimental values with three modes (patient—gene—experimental platform).

Tensor factorization (TF) methods are applied to analyze tensor data just as matrix factorization methods are used for analyzing matrices. TF decomposes a given tensor into factor matrices and a core tensor. A factor matrix encodes latent patterns of each term in the mode and core tensor encodes how patterns of different modes are related. For example, in a 3-order tensor analysis, an input tensor (e.g. PanCan12 tensor) $\mathbf{X}$ is decomposed to a core tensor $\mathbf{G}$ and three factor matrices $A^{(S)}$, $A^{(G)}$ and $A^{(P)}$ (Fig. 1 left). After factorization, one or combinations of the factor matrices and core tensor are used to extract meaningful information.

Applications of TF include anomaly detection from network traffic data (Eliassi-Rad *et al.*, 2006), healthcare monitoring from sensor data (Wang *et al.*, 2017), fraud detection from social network data and biomedical data (Kim *et al.*, 2017). However, tensor factorization methods have not been extensively applied to multi-platform genomic data, mainly due to scalability, missing data problem and interpretability. For example, the PanCan12 dataset, which we test on, forms a 3-order tensor of size $4555 \times 14\,351 \times 5$. If the size of PanCan12 dataset increases (e.g. more patients or platforms), a regular tensor decomposition method will not run due to intermediate data explosion during its calculation (Jeon *et al.*, 2016b).

We have previously addressed the scalability and missing data problems (Choi *et al.*, 2017; Shin *et al.*, 2017) and various ways to exploit prior knowledge to obtain high-quality factorizations (Jeon *et al.*, 2016a). In the case of interpretability, when input data are very sparse and human readable, such as node associations in network tensors, samples of input data can be used as one of the factor matrices resulting in a sparse output that is more interpretable (Lee *et al.*, 2017). However, most scientific data, including PanCan12 data, contain floating point values and are not sparse enough, which makes human interpretation a challenge. This requires a different approach for solving the interpretability problem in the tensor analysis for better discoveries and explanations of latent patterns while preserving the speed and accuracy of factorizations.

Our goal is to devise an interpretable TF method for partially observed tensors exploiting prior knowledge while preserving the accuracy and scalability. Our proposed methods, Silenced-TF (naive) and GIFT (advanced), do this by extending our scalable and accurate tensor decomposition method, P-Tucker (Oh *et al.*, 2018), such that the selected factor matrix preserves and extends the predefined classification of the terms, e.g. gene membership information in functional gene sets.

1.3 Contributions: Our main contributions are as follows

- **Method.** We propose GIFT (Guided and Interpretable Factorization for Tensors) that outputs interpretable (gene) factor matrix by constraining the factor matrices based on prior classification information (functional gene sets).
- **Experiments.** We validate that GIFT is not only interpretable but highly scalable and accurate (Table 1).
- **Discovery.** We apply GIFT to large-scale multi-platform cancer genome analysis using the PanCan12 dataset and show how the method easily and successfully discovers significant relations between patients with gene sets and gene sets to genes.

2 Materials and methods

In this section, we overview our proposed approach, provide details on the dataset used, describe our baseline approach, i.e. P-Tucker, and two prior knowledge constrained methods proposed, i.e. Silenced-TF and GIFT. Preliminaries of a tensor factorization and detailed derivation of the algorithms are provided in the Supplementary Methods. For the readers unfamiliar with tensor analysis, please view the tensor preliminaries in the Supplementary Methods first.

2.1 Overview

We describe P-Tucker (our baseline method with no mask matrices), Silenced-TF (naive interpretable method) and GIFT (advanced interpretable method) in terms of a multi-platform 3-order tensor with (Sample—Gene—Platform) triples with prior information on functional gene sets (All three methods are extendable to general n-order tensors (Supplementary Methods).). In this setting, an input tensor,

$\mathbf{X}$, is decomposed to a core tensor $\mathbf{G}$ and three factor matrices $\mathbf{A}^{(S)}$, $\mathbf{A}^{(G)}$ and $\mathbf{A}^{(P)}$, which we define as (S)ample-, (G)ene- and (P)latform-factor matrices (Fig. 1 left).

Each column in a factor matrix represents a certain latent pattern or concept related to the dimension. For example, a column of $\mathbf{A}^{(S)}$ indicates a subclass of cancers such a type of hereditary breast cancer, and a row specifies the sample. The values of the column indicate weights between the subclass and the sample. Likewise, a significant component of a column of $\mathbf{A}^{(G)}$, which is pre-determined for Silenced-TF and GIFT, corresponds to a gene set and a row corresponds to a gene. A similar explanation can apply for Platform-factor matrix $\mathbf{A}^{(P)}$.

For deriving Silenced-TF and GIFT, we employ prior knowledge in a form of a mask matrix $\mathbf{M}^{(n)}$ representing a membership of pre-determined classification regarding a mode n and utilize the mask matrix to produce an interpretable factor matrix. Produced factor matrix $\mathbf{A}^{(n)}$ has values concentrated on the corresponding unmasked region of the mask matrix. This allows direct mapping of prior knowledge to the latent patterns found in the corresponding column of the factor matrix. Notice that P-Tucker, GIFT and Silenced-TF produce equivalent results if all components of mask matrices are zeros (unmasked).

2.2 Data processing

We use the PanCan12 (Weinstein *et al.*, 2013) and Hallmark gene set data from MSigDB (Liberzon *et al.*, 2015) collections for generating an input tensor and mask matrices, respectively. Table 2 summarizes the data we used in this paper.

2.2.1 Mask matrix

We generate a gene mask matrix $\mathbf{M}^{(G)}$ in a form of (gene—gene set). Each column of mask matrix $\mathbf{M}^{(G)}$ corresponds to a gene set. If a gene i is contained in a gene set j then it is unmasked, i.e. $\mathbf{M}^{(\mathcal{G})}_{ij}$ is set to 0; otherwise, the gene is masked, i.e. set to 1. If no prior-knowledge is known, the elements of mask matrices are all set to zero. In our test, sample mask matrix $\mathbf{M}^{(S)}$ and platform mask matrix $\mathbf{M}^{(P)}$ are filled with zeros.

For the functional gene set, we chose Hallmark collection from the MSigDB (Liberzon *et al.*, 2015) among the various functional gene sets since it has low redundancy and concise mapping to important biological processes. The collection contains 50 independent, refined and concise gene sets that were generated from combining and removing redundancies in various well-known functional gene groups. Thus, there are 50 columns in the generated gene mask matrix, $\mathbf{M}^{(G)}$.

2.2.2 PanCan12 tensor

We transform PanCan12 to a 3-order tensor of size 4555 (sample) × 14 351 (gene) × 5 (platform), where the value of an observed entry indicates the preprocessed experimental value, as follows. Initially, the 4.7 version of the PanCan12 was downloaded from the Sage Bionetworks repository, Synapse (Omberg *et al.*, 2013). The PanCan12 contains multi-platform data with mapped clinical information of patients group into cohorts of twelve cancer type: bladder urothelial carcinoma (BLCA), breast adenocarcinoma (BRCA), colon and rectal carcinoma (COAD, READ), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), acute myeloid leukemia (LAML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous carcinoma (OV) and uterine corpus endometrial carcinoma (UCEC). The five platforms types used are miRNA-seq (MIR), methylation (MET), somatic mutation (MUT), gene expression (GEX) and copy number variation (CNV). After download, probes of each platform were mapped to corresponding gene symbols. Then, samples that have less than two evidence were removed from the dataset. The resulting data for each platform was min-max normalized and was further normalized such that the Frobenius norm became one, i.e. $||A|| \equiv \sqrt{\sum_{i,j} |a_{ij}|^2} = 1$.

Table 1. Characteristic comparisons of GIFT with P-Tucker and Silenced-TF

Method	P-Tucker[a]	Silenced-TF	GIFT
Interpretability	Low	**High**	**High**
Accuracy	**High**	Low	**High**
Scalability	**High**	**High**	**High**

[a]Oh *et al.* (2018).

Table 2. Summary of dataset

Dataset	Order	Size	Observable entries
PanCan12 tensor	3	(4555 × 14 351 × 5)	180M
Sampled-PanCan12	3	(4555 × 14 351 × 5)	36–144M
Mask matrix $\mathbf{M}^{(G)}$	2	(14 351 × 50)	7K

M, million; K, thousand.

2.3 Proposed methods

We first describe our baseline method, P-Tucker, a fast and scalable tensor factorization that does not utilize prior knowledge (gene mask matrix). Then we describe how we have extended P-Tucker method to derive two interpretable factor generation methods: Silenced-TF and GIFT.

2.3.1 Baseline approach: P-Tucker

Our baseline approach, P-Tucker, is a time/memory-optimized tensor factorization algorithm for partially observable tensors (Oh *et al.*, 2018). We have shown in our previous work that P-Tucker outperforms other factorization methods of similar kind (Filipović and Jukić, 2015; Oh *et al.*, 2017; Smith and Karypis, 2017) in terms of scalability and accuracy.

The high scalability and accuracy of P-Tucker come from a novel row-wise update rule of factor matrices derived by computing a gradient of loss function (Supplementary Equation S1) with respect to a given row and setting it to zero for minimizing the loss function. The row-wise update rule is applied to the alternating least squares (ALS) technique that updates a set of parameters, e.g. a factor matrix, while fixing all the others and iterates for all parameter sets. The ALS-based row-wise update allows P-Tucker to process all rows of a factor matrix in parallel, which makes the algorithm highly scalable.

Please refer to Supplementary Methods for full derivations of update rule and proofs of the algorithm described for a general N-order tensor.

2.3.2 Silenced-TF

Although our previous method P-Tucker presents high scalability and accuracy, it produces a dense gene factor matrix with too many genes in each column (latent pattern) having significant factor values. This property makes it difficult to map a single function to each latent pattern found. Our first approach to generate an interpretable factor matrix is a method Silenced-TF that naively silences factor values associated with the non-member of pre-defined gene groups.

In other words, it forces factor values of masked regions to be zeros and updates the rests using the row-wise update rule used for P-Tucker.

Specifically, given a PanCan12 tensor $\mathcal{X} \in \mathbb{R}^{I_S \times I_G \times I_P}$ with observable entries Ω, Silenced-TF of rank (J_S, J_G, J_P) finds a core tensor $\mathcal{G} \in \mathbb{R}^{J_S \times J_G \times J_P}$ and factor matrices $\mathbf{A}^{(S)} \in \mathbb{R}^{I_S \times J_S}, \mathbf{A}^{(G)} \in \mathbb{R}^{I_G \times J_G}, \mathbf{A}^{(P)} \in \mathbb{R}^{I_P \times J_P}$ which minimize the loss function subjected to mask matrices $\mathbf{M}^{(n)} \in \mathbb{R}^{I_n \times J_n}$, where $n \in \{S, G, P\}$ as specified Equation (1).

$$\begin{aligned} &\underset{\mathcal{G}, \mathbf{A}^{(S)}, \mathbf{A}^{(G)}, \mathbf{A}^{(P)}}{\text{minimize}} \mathcal{L}(\mathcal{G}, \mathbf{A}^{(S)}, \mathbf{A}^{(G)}, \mathbf{A}^{(P)}, \mathbf{M}^{(S)}, \mathbf{M}^{(G)}, \mathbf{M}^{(P)}) \\ &= \sum_{\forall \alpha=(i_S,i_G,i_P) \in \Omega} \left(\mathcal{X}_\alpha - \sum_{\forall \beta=(j_S,j_G,j_P) \in \mathcal{G}} \mathcal{G}_\beta \prod_{n \in \{S,G,P\}} a^{(n)}_{i_n j_n} \right)^2 \\ &+ \lambda \left(\sum_{n \in \{S,G,P\}} ||\mathbf{A}^{(n)}||_F^2 \right) \\ &\text{subject to } a^{(n)}_{i_n j_n} = 0 \text{ when } m^{(n)}_{i_n j_n} = 1. \end{aligned} \tag{1}$$

More specifically, given 3-order multi-platform data $\mathcal{X}$ with observable entries Ω, we apply the P-Tucker update rule for a single row $\mathbf{a}^{(n)}_{i_n:}$ in a factor matrix $\mathbf{A}^{(n)}$ by calculating three intermediate data δ, $\mathbf{B}^{(n)}_{i_n}$ and $\mathbf{c}^{(n)}_{i_n:}$, where $n \in \{S, G, P\}$ specifies one of the three modes, i.e. (S)ample, (G)ene, or (P)latform (For easiness of identifying each mode of the tensor, we use first letters of each mode, i.e. $n \in \{S, G, P\}$ for (S)ample, (G)ene and (P)latform. The conventional notation for specifying these modes is ordered indices, i.e. $n = 1, 2, 3$.).

$\delta^{(G)}_{(i_S,i_G,i_P)}$ is a length J_n vector whose jth entry, $\delta^{(G)}_{(i_S,i_G,i_P)}(j)$, is

$$\sum_{\forall (j_S,j,j_P) \in \mathcal{G}} \mathcal{G}_{(j_S,j,j_P)} a^{(S)}_{i_S j_S} a^{(P)}_{i_P j_P}. \tag{2}$$

$\delta^{(S)}_{(i_S,i_G,i_P)}$ and $\delta^{(P)}_{(i_S,i_G,i_P)}$ are defined in the same way.

$\mathbf{B}^{(n)}_{i_n}$ is a $J_n \times J_n$ matrix whose (j_1, j_2)th entry is

$$\sum_{\forall (i_S,i_G,i_P) \in \Omega^{(n)}_{i_n}} \delta^{(n)}_{(i_S,i_G,i_P)}(j_1) \delta^{(n)}_{(i_S,i_P,i_G)}(j_2) \tag{3}$$

and $\mathbf{c}^{(n)}_{i_n:}$ is a length J_n vector whose jth entry is

$$\sum_{\forall (i_S,i_G,i_P) \in \Omega^{(n)}_{i_n}} \mathcal{X}_{(i_S,i_G,i_P)} \delta^{(n)}_{(i_S,i_G,i_P)}(j), \tag{4}$$

where $\Omega^{(n)}_{i_n}$ indicates the subset of Ω whose nth mode's index is i_n.

With the above intermediate data, Silenced-TF, like P-Tucker, updates a row $a^{(n)}_{i_n:}$ by an update rule $\mathbf{c}^{(n)}_{i_n:} \times [\mathbf{B}^{(n)}_{i_n} + \lambda \mathbf{I}_{J_n}]^{-1}$, where $\mathbf{I}_{J_n}$ is a $J_n \times J_n$ identity matrix. The difference in Silenced-TF compared to P-Tucker is an additional step of setting $a^{(n)}_{i_n j_n} = 0$ when $m^{(n)}_{i_n j_n} = 1$ after each row-wise update of $a^{(n)}_{i_n:}$. That is, Silenced-TF only updates an entry in the gene factor matrix when the corresponding masking element is zero; otherwise, Silenced-TF sets the entry to 0.

After updating factor matrices, Silenced-TF (or P-Tucker) calculates reconstruction error by the following rule.

$$\sqrt{\sum_{\forall \alpha=(i_S,i_G,i_P) \in \Omega} \left(\mathcal{X}_\alpha - \sum_{\forall \beta=(j_S,j_G,j_P) \in \mathcal{G}} \mathcal{G}_\beta \prod_{n \in \{S,G,P\}} a^{(n)}_{i_n j_n} \right)^2} \tag{5}$$

When the error converges or the maximum iteration is reached, Silenced-TF (and P-Tucker) terminates the update process.

2.3.3 GIFT

Silenced-TF is based on a simple idea that can be applied to any tensor decomposition methods without modifying the original algorithm a lot. However, it has two weaknesses: low accuracy and inability to discover undefined class components. The reconstruction error of Silenced-TF is much higher than that of P-Tucker due to many zero-value entries in its factor matrices. Regarding the latter weakness, Silenced-TF is able to identify the significance of genes that are included in gene sets but is not capable of finding new genes that show association with the latent function. Hence, to overcome these weaknesses, we propose a more advanced method GIFT which tackles the problem employing selective regularization of factor matrices. The main difference between GIFT and other methods is an existence of mask matrices in the regularization term of the loss function, which allows a soft regularization. The specific loss function of GIFT for a 3-order multi-platform tensor is given by the following Equation (6).

$$\begin{aligned} &\mathcal{L}(\mathcal{G}, \mathbf{A}^{(S)}, \mathbf{A}^{(G)}, \mathbf{A}^{(P)}, \mathbf{M}^{(S)}, \mathbf{M}^{(G)}, \mathbf{M}^{(P)}) \\ &= \sum_{\forall \alpha=(i_S,i_G,i_P) \in \Omega} \left(\mathcal{X}_\alpha - \sum_{\forall \beta=(j_S,j_G,j_P) \in \mathcal{G}} \mathcal{G}_\beta \prod_{n=1}^{3} a^{(n)}_{i_n j_n} \right)^2 \\ &+ \lambda \left(\sum_{n \in \{S,G,P\}} ||\mathbf{M}^{(n)} * \mathbf{A}^{(n)}||^2 \right) \end{aligned} \tag{6}$$

GIFT uses $\mathbf{M}^{(n)} * \mathbf{A}^{(n)}$ instead of just $\mathbf{A}^{(n)}$, where * denotes an element-wise multiplication. Specifically for gene factor matrix, the loss function allows GIFT to focus on learning values of genes in pre-defined functional groups but still allows for non-members of the group to gain factor values if original tensor values are highly associated with the non-member gene. This property increases the accuracy GIFT as well as allowing it to discover new genes that show significant relation to the functional group.

Similar to P-Tucker, the algorithm of GIFT is derived by finding a gradient of the loss function with respect to the given row in a factor matrix and setting it to zero (Supplementary Methods S1.3). The gradient is used to update each row of factor matrices in an ALS fashion, i.e. updating each factor row while fixing all others and iterating over all row in all factors matrices (Fig. 2). Algorithm 1 describes how GIFT updates given factor matrices in detail for the 3-order multi-platform tensor. When GIFT updates a row $a^{(n)}_{i_n:}$ (line 7), it requires a diagonal matrix $\mathbf{D}^{(n)}_{i_n}$ where its (j_n, j_n) entry is $m^{(n)}_{i_n j_n}$ (line 6), while P-Tucker uses an identity matrix $\mathbf{I}_{J_n}$.

Algorithm 1 3-order GIFT

Input: A tensor $\mathcal{X} \in \mathbb{R}^{I_S \times I_G \times I_P}$ with observable entries Ω, mask matrices $\mathbf{M}^{(S)}, \mathbf{M}^{(G)}, \mathbf{M}^{(P)}$, rank (J_S, J_G, J_P), and a regularization parameter λ.
Output: A core tensor $\mathcal{G}$ and factor matrices $\mathbf{A}^{(S)}, \mathbf{A}^{(G)}, \mathbf{A}^{(P)}$.
1: initialize $\mathcal{G}$ and $\mathbf{A}^{(S)}, \mathbf{A}^{(G)}, \mathbf{A}^{(P)}$ randomly
2: **repeat**
3: **for** $n \in S, G, P$ **do**
4: **for** $i_n = 1, \ldots, I_n$ **do**
5: calculate intermediate data δ, $\mathbf{B}^{(n)}_{i_n}$, and $\mathbf{c}^{(n)}_{i_n:}$ by Eq. (2) – (4)
6: calculate $\mathbf{D}_{i_n}$, where its (j_n, j_n)th entry is $\mathbf{M}^{(n)}_{i_n j_n}$
7: update a row $a^{(n)}_{i_n:}$ by $\mathbf{c}^{(n)}_{i_n:} \times [\mathbf{B}^{(n)}_{i_n} + \lambda \mathbf{D}_{i_n}]^{-1}$
8: **end for**
9: **end for**
10: compute reconstruction error by Eq. (5)
11: **until** error converges or exceeds maximum iteration

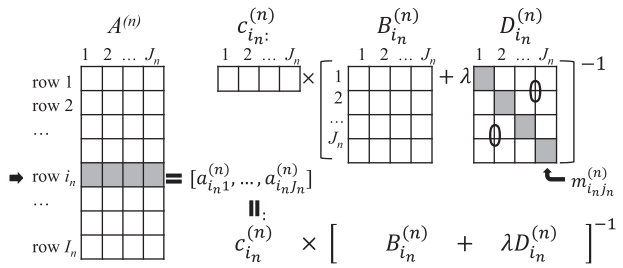

Fig. 2. Row-wise update rule of GIFT

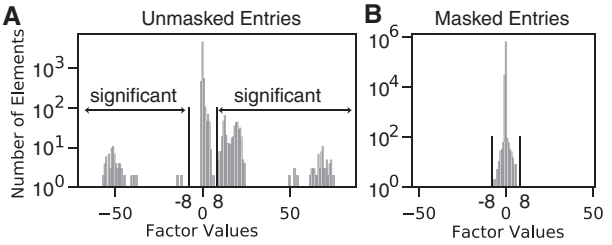

Fig. 3. Distributions of values in a gene factor matrix derived by GIFT ($\lambda = 10$) for unmasked (**A**) and masked (**B**) entries

3 Results

In this section, we describe experimental results of GIFT compared to Silenced-TF and P-Tucker. We aim to answer the following questions.

[Q1] **Interpretability**: How interpretable are factor matrices produced by GIFT and the other methods? (Section 3.2)

[Q2] **Accuracy**: How accurately do GIFT and the other methods factorize a given tensor and predict missing entries of the tensor? (Section 3.3)

[Q3] **Scalability**: How well do GIFT and the other methods scale up with respect to the number of observed entries of a tensor? (Section 3.4)

3.1 Experimental settings

GIFT, Silenced-TF and P-Tucker are implemented in C with OpenMP and Eigen libraries. We run our experiments on a single machine with 20 cores, equipped with an Intel Xeon E5-2630 v4 2.2 GHz CPU and 512GB RAM. We set the default parameters as follows: regularization coefficient $\lambda = 10$ and rank=$(30 \times 50 \times 2)$. Justifications of our parameter selections are summarized in Supplementary Results S2.3. Notice that our convergence criteria include i) when the maximum iteration (20) is reached or ii) when reconstruction error converges (below the threshold; 1%), and we use absolute values of factor matrices for all experiments.

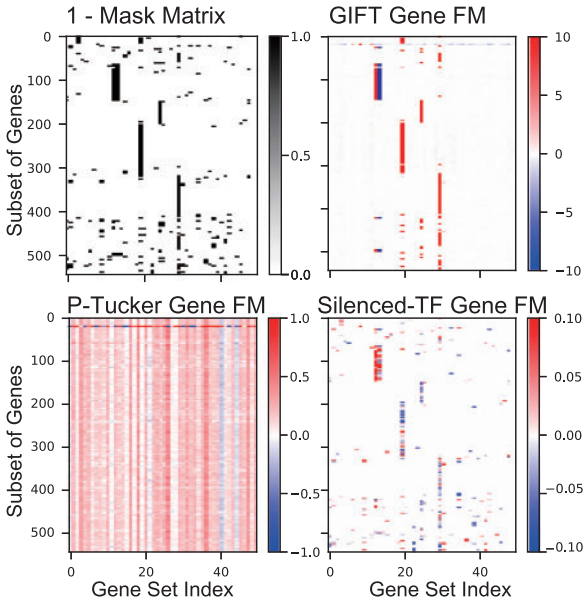

Fig. 4. Mask matrix and gene factor matrices (FM) of GIFT, P-Tucker and Silenced-TF. Subset of genes are shown for better visualization

3.2 Interpretability

We regard a gene factor matrix as interpretable if the set of significant genes that composes each column directly maps to a singular biological function. That is, a gene factor matrix is interpretable if a subset of the genes composing a gene set (unmasked) have significant factor values and a majority of genes that are not in the gene set (masked) have insignificant factor values such that functional information, premapped to each column (gene set), can be used directly in explaining the results. Please view Supplementary Results S2.4 for significant factor threshold selections. Figure 3 shows the distribution of factor values produced by GIFT for unmasked and masked entries.

In this perspective, both Silenced-TF and GIFT had high interpretability (Fig. 4). Compared to a factor matrix produced by P-Tucker, where there was no prior information mapped to a set of significant components in each column, it became easier to interpret factor matrices of Silenced-TF and GIFT where the factor values are concentrated on a predefined gene sets (column components) that already has information mapped to each column. The difference between Silenced-TF and GIFT comes from their ability to explore outside of the pre-determined classification. Our naive model, Silenced-TF, achieved the interpretability by imposing a strict restriction on its factor matrices. Silenced-TF, however, was not able to discover new components outside of pre-defined classification. Our advanced model, GIFT was able to discover outside pre-defined classification by employing a relatively soft restriction on its factors.

Additionally, we define top-K ratios to measure the interpretability of a factor matrix given a mask matrix. A top-K ratio is defined as follows.

$$\text{Top–Kratio R } (0 \leq R \leq 1) = \frac{\text{number of unmasked entries in top-K}}{K} \quad (7)$$

Figure 5 illustrates top-K ratios on varying Ks. P-Tucker showed the worst top-K ratios for all K since it did not distinguish unmasked and masked entries in the calculation. Although Silenced-TF exhibited the highest top-K ratios for all K, Silenced-TF was not able to discover important masked entries which are closely related to unmasked entries. Meanwhile, the top-K ratio of GIFT was the highest until $K \leq 10^2$ and decreased rapidly after $K \geq 10^2$ when the factor values of member genes (unmasked) began saturating and top values began discovering the relevant non-member genes (masked). Overall, Silenced-TF and GIFT provided interpretable factorizations with respect to distributions of values in a factor matrix and top-K ratios.

3.3 Accuracy

We use two evaluation metrics—reconstruction error and test root mean square error (RMSE)—to measure the accuracy. Reconstruction error indicates an accuracy of a factorization as given in Equation (5). Test RMSE implies how accurately a method predicts missing entries of a tensor. To measure test RMSE, we split the PanCan12 tensor into training/test data with a ratio of 9 to 1. As illustrated in Figure 6A and B,

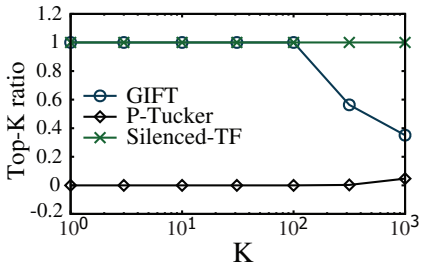

Fig. 5. Top-K ratios based on descending order of absolute factor values

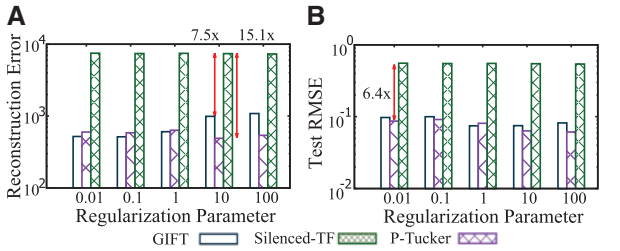

Fig. 6. Performance comparisons of GIFT, Silenced-TF and P-Tucker. (**A**) is a reconstruction error plot. (**B**) is a test RMSE plot

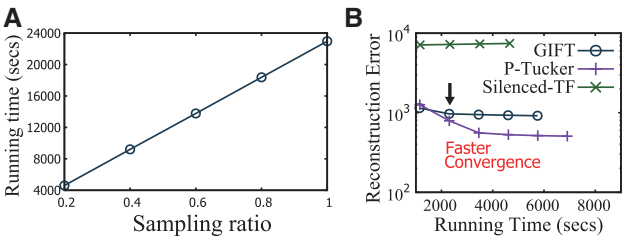

Fig. 7. Convergence and scalability of GIFT. (**A**) Total running time of GIFT with respect to the number of non-zeros. (**B**) GIFT shows faster convergence than P-Tucker and has higher accuracy than Silenced-TF

Silenced-TF exhibits the worst accuracy due to too many zeros in a silenced factor matrix. The reconstruction error and test RMSE of Silenced-TF are $15.1\times$ and $6.4\times$ higher than that of P-Tucker when $\lambda = 10$ and $\lambda = 0.01$, respectively. While P-Tucker shows the best accuracy in most cases, GIFT presents relatively small accuracy loss compared to that of Silenced-TF and converges faster than P-Tucker (Fig. 7B). Furthermore, test RMSE of GIFT is slightly higher or even better than that of P-Tucker. Also, GIFT shows stable accuracy (Supplementary Results S2.2).

3.4 Scalability

Scalability test is performed by varying the number of observable entries by randomly sampling 20, 40, 60, 80 and 100% from the PanCan12 tensor. As shown in Figure 7A, GIFT scales near linearly in terms of the number of observable entries. GIFT also runs with small time and memory overhead for a scaled-up dataset (up to 2 TB; Supplementary Table S6). P-Tucker and Silenced-TF since they present similar scalability to that of GIFT (Supplementary Fig. S5B).

3.5 Empirical validation

We empirically validated GIFT ($\lambda = 10$) by determining whether latent relations found in the patient and gene factor matrices can be mapped to the evidence available in the literature. The latent relations are (cancer—gene sets), (gene sets—genes) and (cancer—genes) found on the PanCan12 dataset.

3.5.1 Patient to gene sets

Given specific patient or group of patients with a cancer type, which gene set is the most relevant? The relevant gene set provides a coarse-grained but holistic view of a patient or a group of patients. We first explain our discovery procedure and validate (patient—gene sets) relations found by GIFT.

We first computed an influence of each gene set on a patient and then calculated the overall influence of each gene set on cancer by aggregating results by the type of cancer. In detail, we consider a row vector $\mathbf{a}^{(1)}_{i:}$ of sample-factor matrix as a latent feature or profile of ith patient, and $\mathbf{G} = (\sum_{i=1}^{I_3} \mathcal{G}_{::i})/I_3$ as a relation between gene sets and columns of the sample-factor matrix. Then, we can regard $\tilde{\mathbf{a}}^{(1)}_{i:} = \mathbf{a}^{(1)}_{i:}\mathbf{G}$ as an influence of each gene set on the ith patient where the jth element of $\tilde{\mathbf{a}}^{(1)}_{i:}$ indicates the influence of jth gene set on the ith patient. We extracted top-5 most important gene sets for each patient by selecting top-5 highest values in $\tilde{\mathbf{a}}^{(1)}_{i:}$. Finally, we counted the frequency of gene sets that appeared in the top-5 gene sets of all patients with a given cancer. We regarded the most frequent gene set as the most relevant one to the given cancer type. The choice $topk = 5$ was chosen based on trial-and-error.

Through the experiment, we found the following latent relationships between gene set mapped functions and cancer types. For breast cancer (BRCA), GIFT considers 'Estrogen response late' and 'Bile acid metabolism' gene sets as related. It is well known that the estrogen plays a key role in the occurrence of breast cancer. The relation to 'Bile acid metabolism' was backed up by Murray *et al.* (1980) where they have shown that patients with breast cancer have significantly low fecal bile acid concentration than that of controlled patients. For ovarian cancer (OV), a relation to the 'Interferon-gamma response' gene set was supported by Wall *et al.* (2003). They showed that interferon-gamma causes apoptosis in human epithelial ovarian cancer. The 'TGF beta signaling' gene set was frequent among many types of cancer including Head and Neck Squamous Cell Carcinoma (HNSC), Lung adenocarcinoma (LUAD), Lung Squamous, Cell Carcinoma (LUSC) and Bladder carcinoma (BLCA). The reason is that the Transforming growth factor-β (TGF-β) gene set is a tumor suppressor which affects many types of human cancers (Kretzschmar, 2000). Additional (cancer—gene sets) relations found are shown in the first and second columns of the Table 3.

3.5.2 Gene sets to genes

Given a gene set, which genes are relevant or irrelevant to provided data? Are there genes not included but related to the gene set according to the data? A significant value in the gene factor matrix indicates that the corresponding gene is highly related to the corresponding gene set. We sorted the genes in each column of the gene factor matrix in descending order by their value and inspected genes with high-absolute factor values for each gene set.

Some of the identified (gene sets—genes) with literature evidence are described in the following and listed in second and third columns of Table 3. GIFT on PanCan12 data identified SKIL gene, known to encodes a protein which antagonizes TGF-β signaling (Tecalco-Cruz *et al.*, 2012), in the 'TGF beta signaling' gene set column to be significant. Likewise, PF4 gene, known as an inhibitor of cell proliferation and angiogenesis (Bikfalvi, 2004), in the 'Angiogenesis' gene set column also had significant factor value; and IRF7 gene, that encodes interferon regulatory factor 7, in the 'Interferon-gamma response' gene set is also identified to be significant.

Table 3. Significant relations found on the PanCan12 dataset via GIFT

Cancer	Gene set	Genes	Evidence
HNSC, LUAD, LUSC, BLCA	TGF beta signaling	SKIL*	Encodes the SNON, negative regulators of **TGF-beta signaling** (Tecalco-Cruz *et al.*, 2012)
		FKBP1A*	Interacts with a type I **TGF-beta receptor**.
		LEFTY2*	Encodes a secreted ligand of the **TGF-beta family** of proteins.
GBM	Angiogenesis	PF4*	Inhibits cell proliferation and **angiogenesis** in vitro and in vivo (Bikfalvi, 2004).
		VCAN*	Encodes a protein involving in celladhesion, and **angiogenesis** (Wight, 2002).
BRCA	Estrogen response late	IL17RB*	Involved in development and progression of **breast cancer** (Alinejad *et al.*, 2017).
		TFF3*	Promotes invasion and migration of **breast cancer** (May and Westley, 2015).
	Bile acid metabolism	APOA1*	**Breast cancer** risk factor (Martin *et al.*, 2015).
OV, UCEC	Interferon-gamma response	IRF7*	Encodes **interferon** regulatory factor 7.
		BST2*	High levels of BST2 have been identified in **ovarian cancer** (Shigematsu *et al.*, 2017).
	Apoptosis	CASP8AP2^{+}	Associated with **apoptosis** of leukemic lymphoblasts (Flotho *et al.*, 2006). Encoded protein plays a regulatory role in Fas-mediated **apoptosis** (Imai *et al.*, 1999).
READ, COAD	Protein secretion	STX7*	Controls vesicle trafficking events involved in **cytokine secretion** (Achuthan *et al.*, 2008).
KIRC, LAML	Mitotic spindle	LATS1*	Binds phosphorylated zyxin and moves it to the **mitotic spindle**

Note: GIFT extracts significant gene sets and notable relations between cancer, gene sets and genes. Evidence column lists supporting evidence for either gene to gene set or gene to cancer relations (*: important gene, $^{+}$: not included in a gene set, but related).

GIFT was also able to identify non-member genes to be related to the gene set. For example, factor value of CASP8AP2 gene was significant in the 'Apoptosis' gene set column and we were able to find literature evidence mapping The CASP8AP2 gene to apoptosis of leukemic lymphoblasts (Flotho *et al.*, 2006).

3.5.3 Cancer to genes findings

Given specific cancer type, which genes affect the cancer type most? We suggest (cancer—genes) relations by combining two relations (cancer—gene sets) and (gene sets—genes) discovered by GIFT.

The first and third columns of Table 3 show (cancer—genes) relations found by GIFT. We regard gene sets in the second column of the table as bridges for (cancer—genes) relations. We deduced the IL17RB and TFF3 genes are significant to breast cancer since the genes are both important for the 'Estrogen response late' gene set and the gene set is the most relevant one to breast cancer. Alinejad *et al.* (2017) showed that IL17RB was crucial in development and progression of breast cancer in effect. Moreover, May and Westley (2015) reveal that the TFF3 gene promoted invasion and migration of breast cancer. GIFT also found that the APOA1 gene in the 'Bile acid metabolism' gene set was highly related to breast cancer. High levels of APOA1 are known to be related to increased breast cancer risk (Martin *et al.*, 2015). In the case of ovarian cancer, GIFT asserts a strong relation to the BST2 gene. High levels of BST2 have been identified in ovarian cancer (Shigematsu *et al.*, 2017).

4 Discussions and conclusion

In this paper, we proposed two scalable and interpretable tensor factorization methods: Silenced-TF and GIFT. The scalability of the two methods come from a parallel computation of factor rows derived from a row-wise update rule. The interpretability of the gene factor matrix is achieved by guiding the factorization to gain values in accordance with the mask matrix that encodes functional gene sets and gene set member information. Our naive model, Silenced-TF, achieves the interpretability by imposing a strict restriction on its factor matrices. Silenced-TF, however, has low accuracy and cannot discover new components outside of pre-defined classification. Our advanced model, GIFT, achieves high accuracy and is able to discover outside of pre-defined classification by employing a relatively soft restriction on its factors.

We applied GIFT to human cancer analytic using the PanCan12 dataset. GIFT was able to find relations between (cancer—gene sets), (gene set—gene) and (cancer—gene) relations, and we were able to find literature evidence to validate their correctness. In finding latent (gene set—gene) relations, GIFT is able to extract out-of-the-box relations, which are not given in prior information.

A notable characteristic of the Silenced-TF and GIFT are their dependencies on the gene sets used in constructing the results. The dependencies of the methods on the gene sets require careful selection of gene sets appropriated for the problem at hand. However, how careful the gene sets were selected, the function of all genes are not yet known, making gene sets incomplete. GIFT is able to learn factor values for nonmembers of the gene sets, due to the penalization scheme adapted for the nonmembers, the factor values learned for the nonmembers tends to have small norms even if the signals from the data are strong. A possible approach to alleviating the inherent incompleteness of gene sets is running GIFT repeatedly and adding nonmembers with relatively high norm values to the new gene set members in the next run of GIFT.

Although GIFT was only applied on a 3-order PanCan12 tensor, it is easily generalized to higher-order tensors as well as larger datasets and various platform data. We believe that GIFT will provide a powerful and extendable tool for large-scale multi-platform genome analysis.

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF-2015R1C1A2A01055739).

Conflict of Interest: none declared.

References

Achuthan,A. *et al*. (2008) Regulation of the endosomal snare protein syntaxin 7 by colony-stimulating factor 1 in macrophages. *Mol. Cell Biol*., **28**, 6149–6159.

Alinejad,V. *et al*. (2017) The role of il17b-il17rb signaling pathway in breast cancer. *Biomed. Pharmacother*., **88**, 795–803.

Anaya,J. *et al*. (2015) A pan-cancer analysis of prognostic genes. *PeerJ*, **3**, e1499.

Bikfalvi,A. (2004) Platelet factor 4: an inhibitor of angiogenesis. *Semin. Thromb. Hemost*., **30**, 379–385.

Choi,D. *et al*. (2017) Fast, accurate, and scalable method for sparse coupled matrix-tensor factorization. *arXiv Preprint arXiv: 1708.08640*.

Eliassi-Rad,T. *et al*. (eds.) (2006) In: *SIGKDD 2016, Philadelphia, PA, USA, August 20–23, 2006*. ACM.

Filipović,M. and Jukić,A. (2015) Tucker factorization with missing data with application to low-n-rank tensor completion. *Multidimensional Syst. Signal Process*., **26**, 677–692.

Flotho,C. *et al*. (2006) Genes contributing to minimal residual disease in childhood acute lymphoblastic leukemia: prognostic significance of casp8ap2. *Blood*, **108**, 1050–1057.

Hoadley,K.A. *et al*. (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, **158**, 929–944.

Hofree,M. *et al*. (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.

Imai,Y. *et al*. (1999) The CED-4-homologous protein FLASH is involved in Fas-mediated activation of caspase-8 during apoptosis. *Nature*, **398**, 777–785.

Jeon,B. *et al*. (2016a) Scout: scalable coupled matrix-tensor factorization-algorithm and discoveries. In: *ICDE 2016*. IEEE, pp. 811–822

Jeon,I. *et al*. (2016b) Mining billion-scale tensors: algorithms and discoveries. *VLDB J*., **25**, 519–544.

Kandoth,C. *et al*. (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.

Kim,S. *et al*. (2015) A mutation profile for top-k patient search exploiting gene-ontology and orthogonal non-negative matrix factorization. *Bioinformatics*, **31**, 3653–3659.

Kim,Y. *et al*. (2017) Discriminative and distinct phenotyping by constrained tensor factorization. *Sci. Rep*., **7**, 1–12.

Koboldt,D.C. *et al*. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Kretzschmar,M. (2000) Transforming growth factor-β and breast cancer: transforming growth factor-β/smad signaling defects and cancer. *Breast Cancer Res*., **2**, 107.

Lee,J. *et al*. (2017) CTD: fast, accurate, and interpretable method for static and dynamic tensor decompositions. *arXiv, Preprint arXiv: 1710.03608*.

Liberzon,A. *et al*. (2015) The molecular signatures database hallmark gene set collection. *Cell Syst*., **1**, 417–425.

Martin,L.J. *et al*. (2015) Serum lipids, lipoproteins, and risk of breast cancer: a nested case–control study using multiple time points. *J. Natl. Cancer Inst*., **107**, djv032.

May,F.E. and Westley,B.R. (2015) Tff3 is a valuable predictive biomarker of endocrine response in metastatic breast cancer. *Endocr. Relat. Cancer*, **22**, 465–479.

Murray,W. *et al*. (1980) Faecal bile acids and clostridia in patients with breast cancer. *Br. J. Cancer*, **42**, 856–860.

Oh,J. *et al*. (2017) S-hot: scalable high-order tucker decomposition. In: WSDM.

Oh,S. *et al*. (2018) Scalable tucker factorization for sparse tensors – algorithms and discoveries. In: *ICDE 2018*, Paris, France.

Omberg,L. *et al*. (2013) Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat. Genet*., **45**, 1121–1126.

Riaz,N. *et al*. (2017) Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat. Commun*., **8**, 857.

Shigematsu,Y. *et al*. (2017) Overexpression of the transmembrane protein bst-2 induces akt and erk phosphorylation in bladder cancer. *Oncol. Lett*., **14**, 999–1004.

Shin,K. *et al*. (2017) Fully scalable methods for distributed tensor factorization. *IEEE TKDE*, **29**, 100–113.

Smith,S. and Karypis,G. (2017) Accelerating the Tucker decomposition with compressed sparse tensors. In: Europar.

Tecalco-Cruz,A.C. *et al*. (2012) Transforming growth factor-β/smad target gene skil is negatively regulated by the transcriptional cofactor complex snon-smad4. *J. Biol. Chem*., **287**, 26764–26776.

Thomas,J. and Sael,L. (2015). Overview of integrative analysis methods for heterogeneous data. In: IEEE BigComp 2015. pp. 266–270

Thomas,J. and Sael,L. (2017) Multi-Kernel LS-SVM based integration bio-clinical data analysis and application to ovarian cancer. *IJDMB*, **19**, 150–167.

Vaske,C.J. *et al*. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.

Wall,L. *et al*. (2003) Ifn-γ induces apoptosis in ovarian cancer cells in vivo and in vitro. *Clin. Cancer Res*., **9**, 2487–2496.

Wang,X. *et al*. (2017) Tensorbeat: tensor decomposition for monitoring multi-person breathing beats with commodity wifi. ACM TIST, **9**.

Weinstein,J.N. *et al*. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet*., **45**, 1113–1120.

Wight,T.N. (2002) Versican: a versatile extracellular matrix proteoglycan in cell biology. *Curr. Opin. Cell Biol*., **14**, 617–623.

Zhu,R. *et al*. (2017) A robust manifold graph regularized nonnegative matrix factorization algorithm for cancer gene clustering. *Molecules*, **22**, 2131.