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

OXFORD

Subject Section

SNeCT: Integrative cancer data analysis via large scale network constrained tensor decomposition

Corresponding Author [1,*], Co-Author [2] and Co-Author [2,*]

[1]Department, Institution, City, Post Code, Country and
[2]Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Network constrained integrative analysis of multiplatform cancer data .
Results: Scalable matrix-tensor coupled Tucker decomposition on PANCAN12 dataset consisting of 5 data types
Availability: The executable and preprocessed data available at http:
Contact: sael@cs.stonybrook.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Integrative analysis of multiple perspective of a patient helps in both cancer stratification and clinical predictions. Stratification help the researchers in understanding and exploring the genomic characteristics in relationship with their current phenotypes and thus to recognize opportunities for clinical improvement on stratified groups of patients. In the perspective of personalized medicine, clinical diagnostics and predictions of individual patient is needed and can be done by searching the integrated profile of patient to existing records. In cancer data analysis, an improved stratification and clinical prediction can be achieved by integrative analysis of the multi-platform data such as copy number alteration (CNA), somatic mutation, gene expression, DNA methylation, and microRNA (miRNA) data. The Genomic Data Commons (GDC) Data Portal https://gdc.cancer.gov/ reports diverse genomic information with paired clinical information.

Analysis of one or few data types may not be sufficient for accurate predict or stratification of disease as they only provide information about part if the patients current biological status. Need for integrative data analysis methods is being recognized, however, due to increased sized of data and limited number of uniform data analysis framework, integrative analysis of multiple data types is still a challenging task. Thomas and Sael (2015) presents two general class of heterogeneous data integration methods, i.e., multiple kernel learning and Bayesian network and showed that integrative analysis of multiple data types or platforms improves stratification and predictions of ovarian cancer Thomas and Sael (2016). Also, many problem specific integrative approaches have been proposed to associate the molecular data with the clinical outcome. These includes a software package implemented in R by Louhimo and Hautaniemi (2011) to show the effect of DNA methylation and copy number alterations in gene expression of several known oncogenes for two cancer type, i.e., glioblastoma and ovarian. Kim *et al.* (2012) proposed a graph based integrated framework that constructs a single graph by determining the optimum linear combination coefficient from the multiple graph obtained using CNA, methylation, miRNA, and gene expression data. Sohn *et al.* (2013) modeled the influence of multi-layered genomic features on gene expression traits by modeling an integrative statistical framework based on a sparse regression. The results showed that using CNA, miRNA, and methylation on gene expression in the predictive power for gene expression level is improved over a single data type based analysis. Schafer *et al.* (2009) approach integrated copy number and gene expression by a modified correlation coefficient and an explorative Wilcoxon test to find DNA regions of abnormalities. The recent work also includes model based prediction of clinical outcomes **?**. Mankoo *et al.* (2011) have applied multivariate Cox Lasso model and median time-to-event prediction algorithm on data set integrated from the four genomic data types (CNA, methylation, miRNA, and gene expression data). Yuan *et al.* (2014) evaluated the predictive power of patient survival and binary clinical outcome using clinical data in combination with a single platform either the somatic copy number alteration, DNA methylation, and mRNA, miRNA or protein expression. They showed slight improvement in some cases when clinical information was combined with one of the molecular. Although this paper showed the predictive power for clinical data in combination with a molecular data, all available molecular data was not used integratively. More recently (The Cancer Genome Atlas Network, 2017)

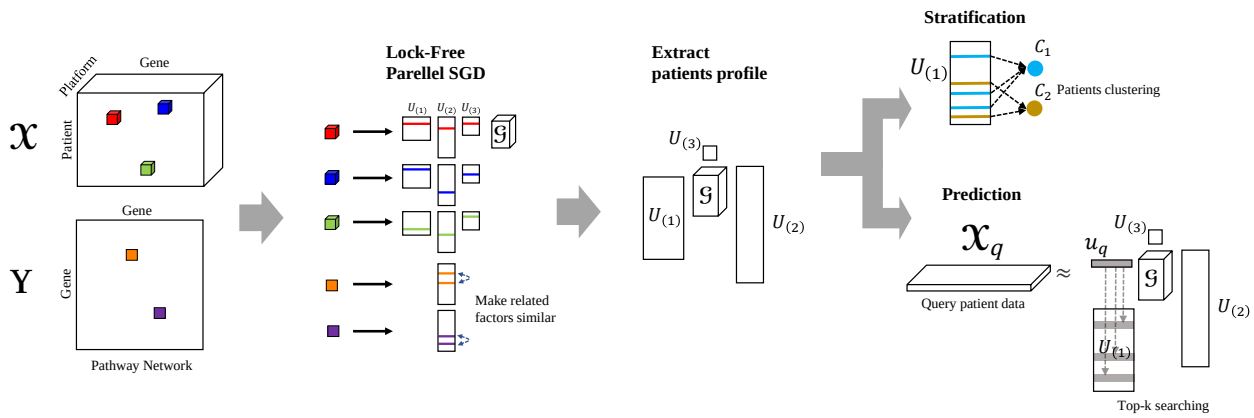


Fig. 1. Overview of the tensor decomposition of SNeCT and validation processes.

The data scalability challenge of in integrative analysis is even more evident in multi-platform data analysis across multiple cancer types. Analysis across multiple cancer types enable us to get a glimpse of extent to which genomic signatures are shared across the different cancers. The biological understanding of similarity and dissimilarity among the different cancers types can enable efficient management of diseases as well as treatment transfers between different cancer types of similar genomic signatures. The work of Hoadley *et al.* (2014) is one the first attempts to utilize multi-platform data of multiple cancers, i.e. the PanCan12 dataset. The PanCan12 data set is created by the Pan-Cancer initiative that compares 12 tumor types profiled by The Cancer Genome Atlas (TCGA) Research Network and includes data from six different platforms (Chang *et al.*, 2013). In there work, integrated subtype classification for all of the tumor samples where performed by first clustering on individual data platforms, and then using the results of single-platform clusters as input to a second-level cluster analysis to form a cluster-of-cluster assignment (COCA). Fully integrative data analysis method would consider and optimize the multi-platform data altogether. In this perspective, the COCA method can be considered as a ensemble method where the input data are varied, rather than a integrative method. Also, it is difficult to utilize the COCA approach for finding similar patients as needed in the clinical predictions given a new patient's data without redoing the analysis over again. Thus, there is a need for multi-platform data analysis method that can scalably stratify multiple cancer types for knowledge discovery and predict clinical outcomes for enabling personalized medicine.

Tensors, i.e., multi-dimensional arrays, are natural representation of multi-platform genomic data (Sael *et al.*, 2015). The core of tensor analysis is tensor decomposition, which can be considered as N-mode singular vector decomposition (SVD). Tensor analysis or tensor mining have been widely applied with success on network traffic (Maruhashi *et al.*, 2011), knowledge bases (Carlson *et al.*, 2010; Nickel *et al.*, 2012; **?**), hyperlinks and anchor texts in the Web graphs (Kolda and Bader, 2006), sensor streams (Sun *et al.*, 2006), and DBLP conference-author-keyword relations (Kolda and Sun, 2008), to name a few. The major challenge of tensor analysis is data scalability as their is a intermediate data explosion problem in the decomposition process even when the input tensor fits into the memory. To address this problem we have previously proposed Hadoop based parallel tensor decomposition method (Jeon *et al.*, 2016*b*,*a*) and approximation method (Shin *et al.*, 2017) that can run on multi-threaded single machines.

In this paper, we propose a tensor-based that enables stratification and clinical prediction of patients utilizing multi-platform data analysis across multiple cancer types. We have shown in our previous works (Kim *et al.*, 2015, 2014) that somatic mutation profiles generated from orthogonal matrix decomposition enables accurate stratification and clinical predictions of each cancer types. We extend this approach to multi-platform multi-cancer data analysis by proposing a scalable network constrained tensor decomposition (SNeCT) method (Figure **??**) and showing that SNeCT can efficiently stratify cancer subtypes and predict clinical outcomes. The contributions of the paper is listed in the following.

- A novel scalable network constrained tensor decomposition (SNeCT) algorithm.
- Generation of multi-platform genomic profile for patients
- Cancer stratification across 12 different cancer types using multi-platform data
- Individualized clinical prediction utilizing genomic profiles of multi-platform data

2 Materials and Methods

Table 1. TCGA PAN Cancer (PanCan12) freeze 4.7 and Synapse repository.

Platform	Input Data Matrix	# of Genes	# of subjects
P1. miRNA	syn2491366		
P2. Methylation	syn2486658		
P3. Somatic copy number (SCNA)	syn1710678		
P4. mRNA	syn1715755		
P5. Somatic mutation	syn1729383		

2.1 Data processing

2.1.1 The PanCan12 dataset

Initially dataset was downloaded from the December 22, 2012 Pan-Cancer-12 data freeze from the Sage Bionetworks repository, Synapse (Omberg *et al.*, 2013). The PanCan12 contains bio-clinical data of patients with one of the twelve cancer type in the following: 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 leukaemia (LAML; conventionally called AML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC),

ovarian serous carcinoma (OV), and uterine corpus endometrial carcinoma (UCEC). Table 1 list the Synapse IDs of the initially download data for each platform used. Then probes of each platform is mapped to a gene symbol. After which subjects and genes that has less then three evidences is removed from the dataset. The resulting data is min-max normalized for each platform and the input tensor is further normalized such that the Frobenius norm, i.e.,$\|A\|_F \equiv \sqrt{\sum_i \sum_j |a_{ij}|^2}$, become one. The cells of resulting 3-mode tensor contain floating point value for $< subject, gene, platform >$ combination as shown in the left of Figure 1. The size of the first mode spanning over subject or the patient index is 4,555; size of the second mode spanning over the genes is 14,351; and the size of the third mode spans over five different platforms.

2.1.2 Pathway data

Batch download of version 8 of bio-network data obtained from the PathwayCommons (Cerami *et al.*, 2011) is used to construct adjacency matrix of gene network for the list of gene consider in the tensor construction. PathwayCommons combines various human gene association information from various bio-network databases. The adjacency matrix contains 665,429 number of association information of 14,351 genes.

2.2 Decomposition and Prediction with SNeCT

2.2.1 Tensor Basics

We describe the basic notations and operations on tensor and its decompositions. Table 2 shows the definitions of symbols used in this paper. A tensor is a multi-dimensional array. An N-dimensional tensor is denoted by $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$. $x_{i_1 i_2 \cdots i_N}$ denotes the $(i_1 i_2 \cdots i_N)$-th element of $\mathcal{X}$. A matrix is denoted by an uppercase bold letter, e.g. $\mathbf{A}$. The i-th row vector of $\mathbf{A}$ is denoted by $\mathbf{a}_i$, a lowercase bold letter, and the (ij)-th entry of $\mathbf{A}$ is denoted by a_{ij}. All tensor and matrix indices are positive integers greater than or equal to 1. The mode-n matrix product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with a matrix $\mathbf{A} \in \mathbb{R}^{I_n \times K}$ is denoted by $\mathcal{X} \times_n \mathbf{A}$ and has the size of $I_1 \times \cdots I_{n-1} \times K \times I_{n+1} \cdots \times I_N$. It is defined element-wise:

$$(\mathcal{X} \times_n \mathbf{A})_{i_1 \ldots i_{n-1} j i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \ldots i_N} a_{j i_n} \tag{1}$$

See Kolda and Bader (2009) for detailed explanations about tensor operations. We focus on 3-D tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ for the following sections since our dataset includes a 3-D tensor as explained in Section 2.1.1.

2.2.2 Higher Order Singular Value Decomposition

Higher order singular value decomposition (HOSVD), also known as Tucker decomposition, is the generalization of singular value decomposition (SVD) which is for matrices (i.e. 2-D tensors). HOSVD decomposes a tensor into a core tensor and orthogonal factor matrices corresponding to each dimensions. Specifically, given a 3-D data tensor $\mathcal{X}$, HOSVD decomposes $\mathcal{X}$ as follows:

$$\mathcal{X} \approx \tilde{\mathcal{X}} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \tag{2}$$

where $\mathcal{G}$ is a core tensor, $\mathbf{U}^{(n)}$ denotes the factor matrices for the n-th dimensions, respectively. More specifically, element-wise formulation of HOSVD is

$$\begin{aligned} x_{i_1 i_2 i_3} \approx \tilde{x}_{i_1 i_2 i_3} &= \mathcal{G} \times_1 \mathbf{u}^{(1)}_{i_1} \times_2 \mathbf{u}^{(2)}_{i_2} \times_3 \mathbf{u}^{(3)}_{i_3} \\ &= \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \sum_{j_3=1}^{J_3} g_{j_1 j_2 j_3} u^{(1)}_{i_1 j_1} u^{(2)}_{i_2 j_2} u^{(3)}_{i_3 j_3} \end{aligned} \tag{3}$$

Table 2. Table of symbols.

Symbol	Definition
$\mathcal{X}$	a tensor (boldface Euler script)
x_{ijk}	(ijk)-th entry of $\mathcal{X}$
$\mathbf{A}$	a matrix (uppercase, bold letter)
$\mathbf{a}_i$	the i-th row vector of $\mathbf{A}$ (lowercase, bold letter)
a_{ij}	(ij)-th entry of $\mathbf{A}$
$\times_n$	n-mode matrix product
$\| \bullet \|$	Frobenius norm
$*$	Hadamard product
$\circ$	Outer product
$\oslash$	Element-wise division
$\Omega_{\mathcal{X}}$	index set of $\mathcal{X}$
$\Omega^{n,i}_{\mathcal{X}}$	subset of $\Omega_{\mathcal{X}}$ having i as the n-th index
I_n	length of n-th dimension of input tensor $\mathcal{X}$
J_n	length of n-th dimension of input tensor $\mathcal{G}$

HOSVD finds the factors by minimizing the following objective function:

$$\begin{aligned} f &= \|\mathcal{X} - \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}\|^2 + \lambda R(\mathcal{G}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}) \\ &= \sum_{(ijk) \in \Omega_{\mathcal{X}}} (x_{ijk} - \mathcal{G} \times_1 \mathbf{u}^{(1)}_i \times_2 \mathbf{u}^{(2)}_j \times_3 \mathbf{u}^{(3)}_k)^2 \\ &+ \lambda R(\mathcal{G}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}) \end{aligned} \tag{4}$$

where $R(\mathcal{G}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)})$ is L_2 regularization term.

2.2.3 HOSVD with network constraint

Consider that a matrix $\mathbf{Y}$ is an adjacency matrix of a graph representing network between entities of a dimension. For our data tensor with dimensions of (subject, gene, platform), $\mathbf{Y}$ is the pathway network graph of genes as explained in Section 2.1.2. $\mathbf{Y}$ informs the similarities between genes, e.g., gene i and gene j are similar if $y_{ij} = 1$. To include the similarity constraint to HOSVD, the network graph $\mathbf{Y}$ acts as a regularization as studied in previous works of Narita *et al.* (2012); Li and Yeung (2009). Specifically, we add the network regularization term $\lambda_g f_g$ to the objective function of Equation (4) where the second dimension is constrained, and λ_g is a constant.

$$f_g := tr(\mathbf{U}^{(2)\mathsf{T}} \mathbf{L} \mathbf{U}^{(2)}) \tag{5}$$

where $\mathbf{L}$ is the Laplacian matrix induced from $\mathbf{Y}$ and $tr(\cdot)$ denotes the trace of a matrix. The Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{Y}$ where $\mathbf{D}$ is a diagonal matrix whose i-th diagonal element $d_{ii} = \sum_{\forall j} y_{ij}$. This regularization helps achieve our goal since following equations hold.

$$\begin{aligned} f_g &= tr(\mathbf{U}^{(2)\mathsf{T}} \mathbf{L} \mathbf{U}^{(2)}) \\ &= \sum_{l=1}^{J_2} \Big[\sum_{(k_1 k_2) \in \Omega_{\mathbf{Y}}} y_{k_1 k_2} (u^{(2)}_{k_1 l} - u^{(2)}_{k_2 l})^2 \Big] \\ &= \sum_{(k_1 k_2) \in \Omega_{\mathbf{Y}}} y_{k_1 k_2} \|\mathbf{u}^{(2)}_{k_1} - \mathbf{u}^{(2)}_{k_2}\|^2 \end{aligned} \tag{6}$$

Minimizing f_g leads to the result that $\mathbf{u}^{(2)}_{j_1}$ and $\mathbf{u}^{(2)}_{j_2}$ have similar values when there is an edge between gene j_1 and gene j_2 in the graph $\mathbf{Y}$.

We present a multi-core algorithm to minimize the objective function $f_{opt} = 12f + \frac{1}{2}\lambda_g f_g$ and factorize the given tensor $\mathcal{X}$ into HOSVD form. SNeCT adopts parallel stochastic gradient descent (SGD) optimization

Algorithm 1 SNeCT

Require: Input data: tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, network matrix $\mathbf{Y} \in \mathbb{R}^{I_c \times K}$, number of parallel cores P, and network-constrained mode c
Hyperparameters: core size $(J_1, J_2, \cdots, J_N)$, learning rate η, regularization factors λ and λ_g
Ensure: Core tensor $\mathcal{G} \in \mathbb{R}^{J_1 \times J_2 \times \cdots \times J_N}$, factor matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \cdots, \mathbf{U}^{(N)}$
1: Initialize $\mathcal{G}, \mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times J_n}$ for $n = 1, 2, \cdots, N$ randomly
2: **repeat**
3: **for** $\forall x_{i_1 i_2 \cdots i_N} \in \mathcal{X}$, $\forall y_{k_1 k_2} \in \mathbf{Y}$ in random order **do in parallel**
4: **if** $x_{i_1 i_2 \cdots i_N} \in \mathcal{X}$ is picked **then**
5: Cache intermediate data tensor: $\mathcal{D} \leftarrow \mathcal{G} * (\mathbf{u}^{(1)}_{i_1} \circ \mathbf{u}^{(2)}_{i_2} \circ \cdots \circ \mathbf{u}^{(N)}_{i_N})$
6: $\tilde{x}_{i_1 i_2 \cdots i_N} \leftarrow$ sum of all elements of $\mathcal{D}$
7: Update corresponding factor rows: $\mathbf{u}^{(n)}_{i_n} \leftarrow \mathbf{u}^{(n)}_{i_n} - \eta\big((\tilde{x}_{i_1 i_2 \cdots i_N} - x_{i_1 i_2 \cdots i_N}) \cdot Collapse(\mathcal{D}, n) + \frac{\lambda}{|\Omega^{n,i_n}_{\mathcal{X}}|}\mathbf{u}^{(n)}_{i_n}\big)$, (for $n = 1, 2, \cdots, N$)
8: Update core tensor: $\mathcal{G} \leftarrow \mathcal{G} - \eta P\big((\tilde{x}_{i_1 i_2 \cdots i_N} - x_{i_1 i_2 \cdots i_N}) \cdot \mathcal{D} \oslash \mathcal{G} + \frac{\lambda}{|\Omega_{\mathcal{X}}|}\mathcal{G}\big)$, (executed by only one core)
9: **end if**
10: **if** $y_{k_1 k_2} \in \mathbf{Y}$ is picked **then**
11: Update network-constrained factors: $\mathbf{u}^{(c)}_{k_1} \leftarrow \mathbf{u}^{(c)}_{k_1} - \eta\lambda_g y_{k_1 k_2}(\mathbf{u}^{(c)}_{k_1} - \mathbf{u}^{(c)}_{k_2})$, $\mathbf{u}^{(c)}_{k_2} \leftarrow \mathbf{u}^{(c)}_{k_2} - \eta\lambda_g y_{k_1 k_2}(\mathbf{u}^{(c)}_{k_2} - \mathbf{u}^{(c)}_{k_1})$
12: **end if**
13: **end for**
14: **until** convergence conditions are satisfied
15: $\mathbf{Q}^{(n)}, \mathbf{R}^{(n)} \leftarrow$ QR decomposition of $\mathbf{U}^{(n)}$, $\mathbf{U}^{(n)} \leftarrow \mathbf{Q}^{(n)}$, $\mathcal{G} \leftarrow \mathcal{G} \times_n \mathbf{R}^{(n)}$, (for $n = 1, 2, \cdots, N$)
16: **return** $\mathcal{G}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \cdots, \mathbf{U}^{(N)}$

technique thus our method is highly memory-efficient and scalable to large datasets and multiple cores. We rewrite f so that it forms an SGD-amenable form.

$$f = \sum_{(i_1 i_2 i_3) \in \Omega_{\mathcal{X}}} \Big[\big(x_{i_1 i_2 i_3} - \tilde{x}_{i_1 i_2 i_3}\big)^2 + \frac{\lambda}{|\Omega_{\mathcal{X}}|}\|\mathcal{G}\|^2 + \lambda \sum_{n=1}^{3} \frac{\|\mathbf{u}^{(n)}_{i_n}\|^2}{|\Omega^{n,i_n}_{\mathcal{X}}|}\Big]$$

Note that f_g already has the SGD-amenable form. Now, gradients of f_{opt} with respect to factors for a given data point $x_{\alpha=(i_1 i_2 i_3)}$ or $y_{\beta=(k_1 k_2)}$ are calculated as follows:

$$\begin{aligned}
\left.\frac{\partial f_{opt}}{\partial \mathbf{u}^{(1)}_{i_1}}\right|_{\alpha} &= -\big(x_\alpha - \tilde{x}_\alpha\big)\big[\mathcal{G} \times_2 \mathbf{u}^{(2)}_{(i_2)} \times_3 \mathbf{u}^{(3)}_{(i_3)}\big] + \frac{\lambda}{|\Omega^{1,i_1}_{\mathcal{X}}|}\mathbf{u}^{(1)}_{i_1} \\
\left.\frac{\partial f_{opt}}{\partial \mathcal{G}}\right|_{\alpha} &= -\big(x_\alpha - \tilde{x}_\alpha\big) \times_1 \mathbf{u}^{(1)\mathsf{T}}_{i_1} \times_2 \mathbf{u}^{(2)\mathsf{T}}_{i_2} \times_3 \mathbf{u}^{(3)\mathsf{T}}_{i_3} + \frac{\lambda}{|\Omega_{\mathcal{X}}|}\mathcal{G} \\
\left.\frac{\partial f_{opt}}{\partial \mathbf{u}^{(2)}_{k_1}}\right|_{\beta} &= \lambda_g y_\beta(\mathbf{u}^{(2)}_{k_1} - \mathbf{u}^{(2)}_{k_2})
\end{aligned} \tag{7}$$

$\left.\frac{\partial f_{opt}}{\partial \mathbf{u}^{(2)}_{i_2}}\right|_{\alpha}$, $\left.\frac{\partial f_{opt}}{\partial \mathbf{u}^{(3)}_{i_3}}\right|_{\alpha}$, and $\left.\frac{\partial f_{opt}}{\partial \mathbf{u}^{(2)}_{k_2}}\right|_{\beta}$ are calculated symmetrically as the above equations. The above equations are naturally generalized to mode-N tensors.

2.2.4 SNeCT algorithm

SNeCT optimizes the objective function (4) by parallel SGD update. Algorithm 1 shows detailed procedures of decomposition of a general N-mode tensor $\mathcal{X}$ and network constraint $\mathbf{Y}$ which represents the similarity of c-th mode entities.

In the beginning, SNeCT initializes $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, $\cdots$, $\mathbf{U}^{(N)}$, and $\mathcal{G}$ randomly (line 1 of Algorithm 1). The outer loop (lines 2-14) repeats until the factors converge. In the inner loop (lines 3-13), SNeCT conducts parallel updates of factor rows corresponding to each data point $x_{i_1 i_2 i_3}$ or $y_{j_1 j_2}$ in random order. When calculating gradients with respect to factor rows and core tensor, it takes excessive time to calculate $\tilde{x}_{alpha} = \mathcal{G} \times_1 \mathbf{u}^{(1)}_{i_1} \times_2 \cdots \times_n \mathbf{u}^{(N)}_{i_N}$ and tensor-matrix products for $\frac{\partial f_{opt}}{\partial \mathbf{u}^{(n)}_{i_n}}$ every time when they are needed. SNeCT reduces the time cost efficiently by caching intermediate data tensor $\mathcal{D}$ (line 5). See section 3.5 of Choi *et al.* (2017) for detailed approach of utilizing intermediate data to reduce time cost. $Collapse(\mathcal{D}, n)$ operator (line 7) outputs a vector with length of i_n which contains the sum of k-th slice of D over n-th mode as its k-th element. In line 8, element-wise division operator $\oslash$ is used to efficiently calculate core tensor gradient.

Table 3. Comparison of time complexity (per iteration) and memory usage of SNeCT with existing network-regularized HOSVD algorithm of Narita et al. (2012). SNeCT shows lower time complexity and memory usage. For simplicity, we assume that data tensor $\mathcal{X}$ has order of N, all modes are of size I, of rank J, and one mode has network constraint. P is the number of parallel cores.

	Time complexity (per iter.)	**Memory usage**
SNeCT	$\mathcal{O}(\|\Omega_{\mathcal{X}}\|J^N N/P + \|\Omega_{\mathbf{Y}}\|J/P)$	$\mathcal{O}(J^N P)$
Narita *et al.* (2012)	$\mathcal{O}(\|\Omega_{\mathcal{X}}\|J^N N^2 + \|\Omega_{\mathbf{Y}}\|J)$	$\mathcal{O}(J^{N-1} I)$

There are possible conflicts between the parallel updates since a factor row or core tensor might be accessed by multiple update attempts. However, we apply lock-free parallel update scheme (Recht *et al.*, 2011; Choi *et al.*, 2017) and remove frequent conflicts by updating core tensor using only one core (line 6) thus SNeCT guarantees near-linear convergence to a local optimum.

There is an existing work for network-regularized tensor decomposition of Narita *et al.* (2012) which follows gradient descent approach which is hardly parellelizable and takes high memory requirement due to the batch calculation of matrices for gradients. Table 3 summarizes the comparison between SNeCT and Narita *et al.* (2012). SNeCT outperforms the existing method in terms of both time complexity and memory usage. SNeCT achieves low time complexity by direct parallelization of SGD updates while caching intermediate data tensor and low memory usage by avoiding batch matrix calculation used by Narita *et al.* (2012).

2.2.5 Stratification and top-k search using SNeCT

Hyperparameters for the decomposition determined by the best resulted on validation set (see Section 3.1). After decomposition of $\mathcal{X}$, we get patients factor matrix $\mathbf{U}^{(1)}$. We assign patients into clusters by using k-means clustering algorithm for factor row vectors. Several distance

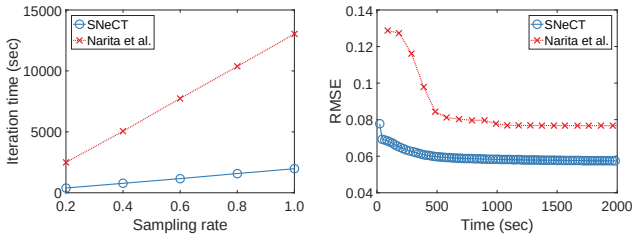
Table 4. 12 pathological disease types assigned to clusters of profiles factorized from tensor factorization.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	Total
BLCA	16	32	2	19	0	22	3	0	0	0	32	0	0	126
BRCA	17	3	600	172	1	70	0	0	0	0	26	0	0	889
COAD	4	0	2	2	0	91	317	0	0	0	1	2	0	419
GBM	4	1	1	2	3	7	0	0	248	0	1	0	0	267
HNSC	0	242	1	6	0	1	0	0	0	0	60	0	0	310
KIRC	14	1	1	0	471	4	0	0	1	0	6	0	0	498
LAML	0	0	0	0	0	9	0	0	0	188	0	0	0	197
LUAD	302	2	2	7	1	12	0	0	0	0	29	0	0	357
LUSC	26	32	0	29	0	7	0	0	0	0	246	0	0	340
OV	0	0	1	3	0	1	1	348	0	0	0	0	131	485
READ	1	1	0	5	0	9	145	0	0	0	1	1	0	163
UCEC	3	1	3	117	1	348	1	0	0	0	10	13	2	499
Total	387	315	613	362	477	581	467	348	249	188	412	17	134	4550

measures (euclidean, cosine and mahalanobis) for clustering have been tested. However, there were no significant differences in the results.

When a new query patient q arrives with data $\mathcal{X}_q$ where $\mathcal{X}_q$ is a tensor representing patient profile of q, we aim to find the HOSVD factor for the patient using the pre-calculated factor matrices and core tensor. Thus we compare the factors of other patients which are encoded in the patient factor matrix $\mathbf{U}^{(1)}$ with the calculated patient factor. We solve the equation $\mathbf{u}_q = \arg\min_{\mathbf{u}} \|\mathcal{X}_q - \mathcal{G} \times_1 \mathbf{u} \times_2 \mathbf{V} \times_3 \mathbf{W}\|$ with our proposed parallel HOSVD algorithm while fixing parameters other than $\mathbf{u}$. $\mathbf{u}_q$ is used to seek top-k similar patients with the given query patient by calculating the distance between the query factor and patient factors to predict the clinical features.

3 Results



(a) Running time (b) Convergence

Fig. 2. Running time and convergence of SNeCT and Narita et al. (2012).

3.1 Performance of SNeCT

We experimentally evaluate SNeCT by measuring its accuracy and speed of decomposition and comparing them with the existing method. First, we compare running time (for one iteration) of SNeCT with its competitor, Narita *et al.* (2012). We create sampled datasets by randomly selecting part of patients with certain ratio to verify the efficiency of SNeCT for 'big data' scenario. Figure 2(a) shows the running time of two methods for one iteration. Both methods scale linearly as the sampling rate (number of patients) increases. SNeCT takes less time than its competitor. That is, for original dataset, running time of SNeCT is 1974s, 6.6× faster than that of Narita *et al.* (2012), 13036s. Not only the running time but also the accuracy of decomposition is critical part of the performance of our method. A factored result can intuitively be evaluated by how well the factored components can reconstruct back the validation tensor. We use the Root Mean Squared Error (RMSE) as measure for reconstruction error. We use a randomly sampled tensor, 10% of patients and 10% of genes to evaluate convergence property of two methods. Figure 2(b) shows the convergence property of two methods. A lower RMSE for the same time means the faster convergence. SNeCT converges to a local minimum within much fewer iterations than Narita *et al.* (2012) while Narita *et al.* (2012) fails to find a proper optimum.

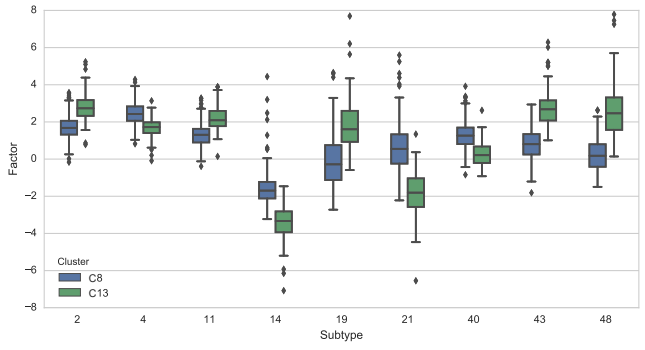
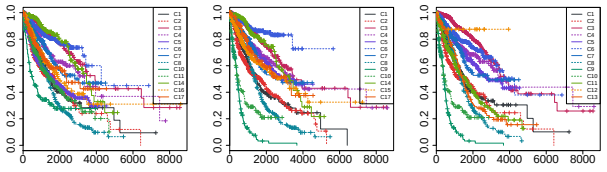


Fig. 3. Distinguishing two clusters for OV.



(a) $\lambda_g = 0$ (b) $\lambda_g = 0.1$ (c) $\lambda_g = 1$

Fig. 4. Predicted survival curves for COCA-clustered patients. x-axis is survival time (day) and y-axis is survival rate.

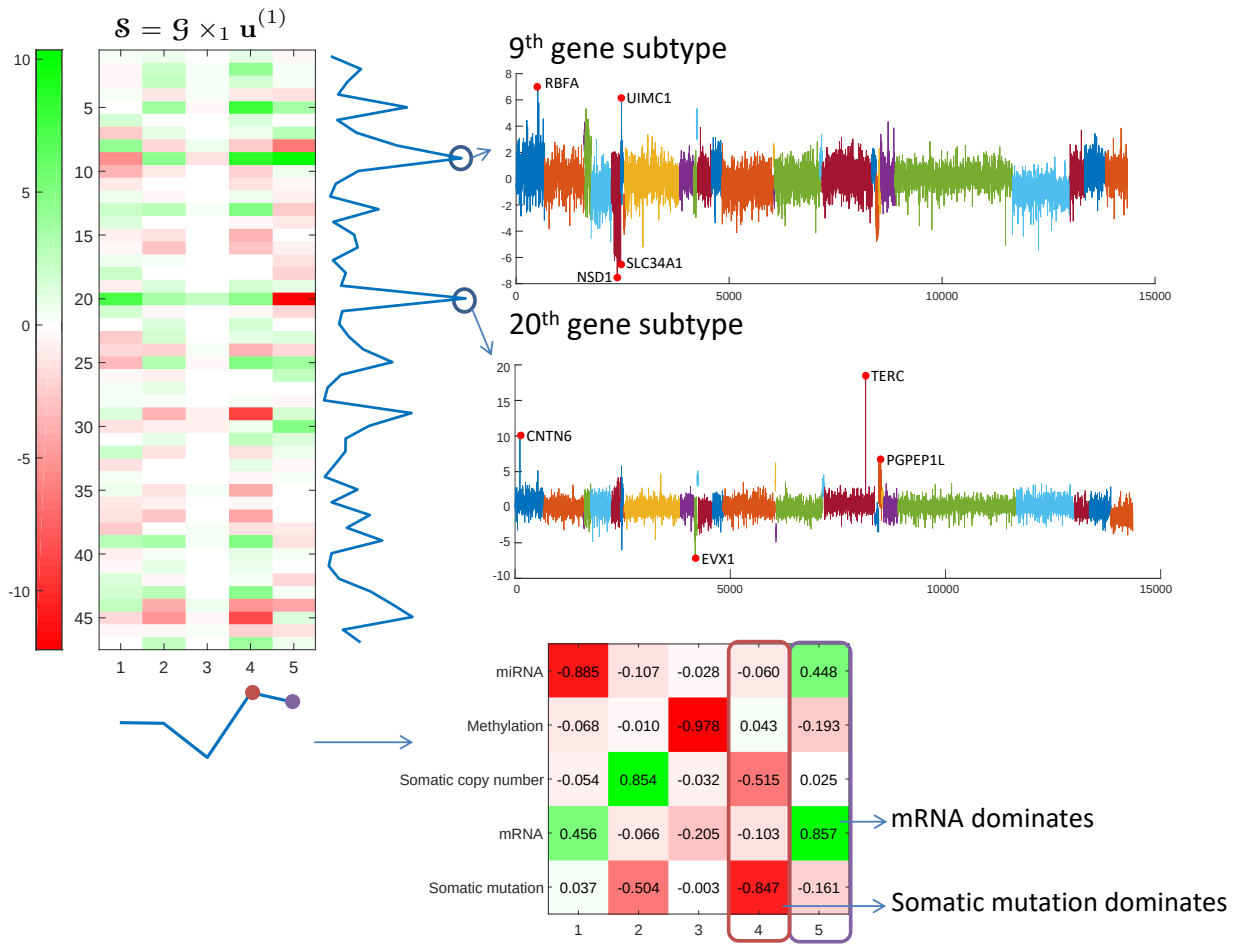


Fig. 5. Personalized gene analysis procedure.

3.2 Stratification

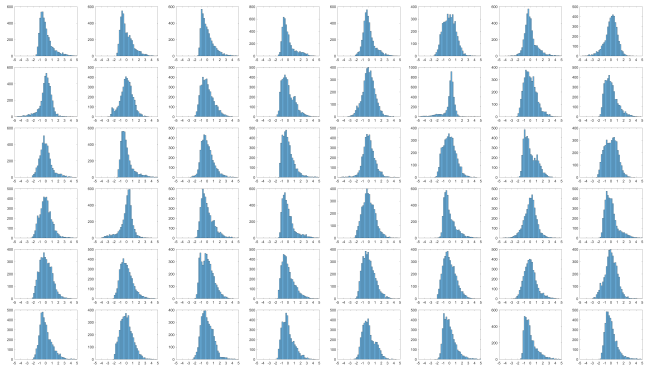


Fig. 6. Gene factor distribution.

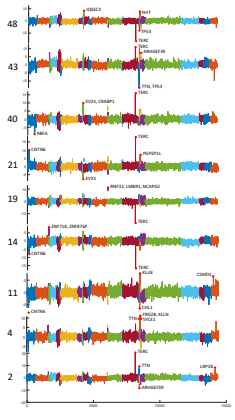


Fig. 7. Subtypes distinguishing OV clusters.

3.2.1 Cluster of cluster assignment

We perform COCA (cluster of cluster assignment) analysis with k-means clustering algorithm for patient profiles. First, we decompose the data tensor with core size $[J_1, J_2, J_3]$ = [78, 48, 5], the best parameters searched with a small validation tensor. We use the ground truth data of patients on which cancer they have among 12 cancer types, 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 leukaemia

(LAML; conventionally called AML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous carcinoma (OV), and uterine corpus endometrial carcinoma (UCEC).

We stratified patient profiles into 13 clusters as suggested by Hoadley *et al.* (2014). Table 5 shows the number of assigned patients for each sub-clusters.

3.2.2 Survival analysis

We performed survival analysis for each COCA subtype we acquired from Section 3.4.1, using the Cox proportional hazards regression model in the R survival package. We use right-censored survival data for 4511 patients whose survival information is known: days to death for dead patients, and days to last contact for alive patients as right-censored data. To see how the network constraint affects decomposition result, we impose three different level of network regularization, λ_g: 0 (not constrained), 0.1, and 1.

The log-rank statistics are computed to compare how well the survival curves are distinguished from each other. the log-rank statistic value when $\lambda_g = 0$ is 409. When $\lambda_g = 0.1$ and 1, the log-rank statistic value are 1151 and 1185, respectively. Higher log-rank statistic means more statistical significance. Based on the calculated value, constraint with $\lambda_g = 1$ has the highest statistic and show most statistical significance for distinguishing patients. When network-constraint is applied ($\lambda_g = 0.1, 1$), SNeCT apparently finds better clusters which distinguishes patients' survival time. For example, seeing Figure 9(b) and (c), SNeCT finds clusters of patients whose survival expectancy notably deviate from most of other patients. Cluster C9, GBM-dominant cluster, has shorter survival time than others. Cluster C12 for $\lambda_g = 1$ mostly contains patients reported as alive.

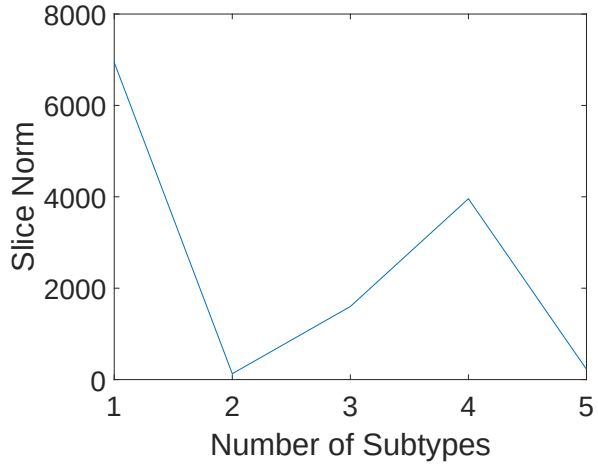


Fig. 8. Norm of slices of $\mathcal{G} \times_3 \mathbf{U}^{(3)}$.

3.2.3 Subtype analysis

Personalized subtype analysis For patient i, SNeCT gets patient profile $\mathbf{u}_i^{(1)}$. We calculate the subtype matrix $\mathcal{S} = \mathcal{G} \times_1 \mathbf{u}_i^{(1)}$ ($\in \mathbb{R}^{J_2 \times J_3}$). $\mathcal{S}$ provides a personalized weight information of subtypes for the gene and the platform modes. Each row of $\mathcal{S}$ represents a subtype for gene mode, and each column represents a platform subtype. By examining the high norm values of rows or columns, we find out influential gene or platform for the patient i. We provide further description for the personalized subtype analysis with an example. Figure 5 shows the personalized subtype map $\mathcal{S}$ of a patient A0UV. normalized norms of rows (gene subtypes) have several pick values. 9th and 20th gene subtypes are the most dominant subtypes. Now we examine the two normalized subtype columns for the gene factor matrix $\mathbf{U}^{(2)}$. Genes with top-4 peak values are marked. Next, we perform platform analysis for the sample patient. The 4th and 5th subtypes are dominant in terms of column norm. Examining $\mathbf{U}^{(3)}$, somatic mutation and mRNA dominate on the 4th, and 5th subtypes, respectively. Thus, we conclude that the gene set {TERC, CNTN6, EVX1, PGPEP1L, NSD1, RBFA, SLC34A1, UIMC1}, and the platform set {somatic mutation, mRNA} are most influential on the mutation data of patient A0UV. The above description only provides the schemes while more quantified comparison is available by multiplying the norm and factor values.

Platform factor analysis Beside the personalized analysis, SNeCT provides whole factor analysis. Here we analyze the degree of influence for each platform (Table 1) on the entire given data. Norm of a slice of $\mathcal{G} \times_3 \mathbf{U}^{(3)}$ corresponding to each platform is shown in Figure 11. Note that P1 (miRNA) is the most dominant, and P4 (mRNA) is the second dominant platform along our entire dataset.

3.3 Prediction

3.3.1 Prediction accuracy of top-k search

SNeCT finds the query profile vector $\mathbf{u}_q$ when query data for a new patient arrives. To simulate the top-k retrieval situation, we select 10% of observed patients in the data as test query set. We assume that each patient in the test set arrives as a query and calculate the top-10 nearest patients for the query patient.

Using BRCA and GBM data, 84.1% and 84.6% of patients have same value for 'neoplasm cancer status 'as their top-1 similar patient, respectively. On average, 71.8% of the patients have value for the clinical feature with the top-1 similar patient. The R-preicision value of BRCA data for predicting estrogen receptor is 80.6%.

3.3.2 Search speed

SNeCT finds the factors for a query with high speed of parallel SGD. Naive search using the query tensor and raw data takes 85560ms, and SNeCT takes 550ms on average.

3.4 Stratification

3.4.1 Cluster assignment

We perform cluster analysis with k-means clustering algorithm on patient profiles. To generate patient profiles, we decompose the data tensor as shown in Fig. 1 with core size [J_1, J_2, J_3] = [78, 48, 5], where the best core size was searched with a small validation tensor and graph constraint is set to $\lambda = 1.0$. After decomposition, the rows of patient factor matrix $U_(1)$ is used as patient profiles. To find the cluster size we compute the gap statistics introduced by Tibshirani *et al.* (2001). The gap statistics of the cluster stabilizes after cluster size of 9 as shown in Fig. S1. For convince of comparison, we stratified patient profiles into 13 clusters as suggested in Hoadley *et al.* (2014). Table 5 shows the number of assigned patients of twelve cancer types for each sub-clusters.

3.4.2 Cluster analysis

Cluster result of the patient factor matrix, $U_(2)$, correlates with with tissue of origin which is similar to observation made by Hoadley *et al.* (2014). Five clusters, C5-KIRC, C8-OV-1, C9-GBM, C10-LAML, and C13-OV-2 each dominantly contains cancer samples of single tissue of origin. C5-KIRC contains 471 patients classified as KIRC and 6 other patients classified to other cancer types. C8-OV-1 contains 348 cases of OV patients with no other cancer type. C9-GBM contains 248 cases of GBM patients with only one of KIRC patient. C10-LAML contains 188 cases of LAML with no other cancer patients included. C12-UCEC-small, a small cluster, contains 13 UCEC with 4 other cases. Finally, C13-OV-2 contains 131 cases of OV patient with only three other cancer types.

Table 5. 12 pathological disease types assigned to clusters of profiles factorized from tensor factorization with graph constraint of $\lambda = 1.0$.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	Total
BLCA	16	32	2	19	0	22	3	0	0	0	32	0	0	126
BRCA	17	3	600	172	1	70	0	0	0	0	26	0	0	889
COAD	4	0	2	2	0	91	317	0	0	0	1	2	0	419
GBM	4	1	1	2	3	7	0	0	248	0	1	0	0	267
HNSC	0	242	1	6	0	1	0	0	0	0	60	0	0	310
KIRC	14	1	1	0	471	4	0	0	1	0	6	0	0	498
LAML	0	0	0	0	0	9	0	0	0	188	0	0	0	197
LUAD	302	2	2	7	1	12	0	0	0	0	29	0	0	357
LUSC	26	32	0	29	0	7	0	0	0	0	246	0	0	340
OV	0	0	1	3	0	1	1	348	0	0	0	0	131	485
READ	1	1	0	5	0	9	145	0	0	0	1	1	0	163
UCEC	3	1	3	117	1	348	1	0	0	0	10	13	2	499
Total	387	315	613	362	477	581	467	348	249	188	412	17	134	4550

Patients in C1-LUAD-enriched cluster includes 302 out of 357 LUAD patients with relatively good prognosis, that is neoplasm cancer status is tumor free for 186 out of all 217 tumor free cases with precision of 0.73 (recall of 0.85).

BRCA: C3: Patients in C3-BRCA/Luminal cluster group 600 BRCA cases with 13 other cancer types. The BRCA patients have positive estrogen receptor status (precision of 0.95) and contains 8 out of 9 metastatic cases. progesterone receptor status is mostly positive with precision of 0.83. HER2 status is mixed tending to have more negative status with precision of 0.73. Contains 34 out of 43 cases with know other malignancy histological type. Which tells us that C3 groups patients with either Luminal A and Luminal B molecular subtypes of breast cancer.

C4: Second C4-BRCA/UCEC-enriched Contains 172 BRCA, 117 UCEC, 29 LUSC, 19 BLCA cases. BRCA: her2 status is mostly negative (101 negative - precision 0.68, 23 positive, 23 equivocal); icd10 of c50.9 (171/880) and c50.8 (1/1); progesterone receptor status is mostly negative (131 negative, 32 positive, 1 intermediate); various histological types however containing all 5 of medullary carcinoma cases; estrogen receptor status is more negative (127 negative - precision 0.77, 38 positive); - normal like (?) UCEC: contains more serous adenocarcinoma (69 out of 115 possible) then the more common endometrioid carcinoma (45 out of 45+333); neoplasm cancer status with with tumor for 29 (74 possible) and tumor free is 77 (77+316 possible); grade of the tumor tends to be high with 7 high grade (out of 11) and 100 of grade 3; contains 12 out of 30 cases with known other maligmancy histological type - bad prognosis(?) LUSC: no strong tendency was found. various smoking history and histology and other labels. (check if mixed tumor) BLCA: 18 of 19 has icd o 3 histology 8120/3 and one out of one 8070/3; all 19 has histological type of muscle invasive urothelial carcinoma (pt2 or above); contains more non-papillary diagnosed cases (16 non-papillary vs 3 papillary); all 19 has high neoplasm histologic grade; 7 of which is know to have other malignancy histological type associated; - bad prognosis (?)

Patients in C6-BRCA/UCEC-enriched: contains 22 BLCA, 70 BRCA, 90 COAD, 348 UCEC cases out of 581 cases. C6 contains 22 BLCA patients contain 15 of icd 10 code of c67.9; icd-o 3 histology 19 of $8120/3$ and 3 of $8130/3$; with no 70 BRCA patients in C6 has is mostly negative HER2 status (38 negative, 8 positive, 9 other) mixed positive and negative status for both progesterone receptor estrogen receptor. 90 COAD: contains all three cases of abnormal braf gene analysis result; 348 UCEC- contains all 3 cases of icd 10 c54.9 and c54.3 in addition to 345 c54.1 cases. contains large portion of icd o 3 histology $8380/3$ (precision of 0.89 - 312 out of 312+63 8380/3 cases) neoplasm cancer status is mostly tumor free (precision of 0.88 - 292 out of 292+101 cases) large portion of histological grade of 1 (87 over 87+4) grade 2 (93 over 93+13) and approximately half of grade 3 and little of high grade (3 out of 3+8)

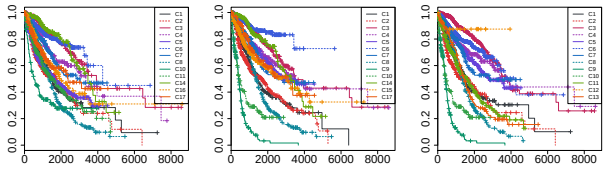
C7-COAD/READ-enriched: contains 317 COAD 145 READ and 5 other cases. COAD, unlike COAD group in C6, C7 contains 22 (precision 1 and recall of 0.88) normal braf gene analysis results. READ subcluster contain 145 out of 163 READ cases.

Patients in C2-HNSC-enriched-squamous-like cluster contains squamous-like BLCA (32), HNSC (242), and LUSC (32) mostly male (228/315) patients. BLCA contains 26 patients diagnosed as non-papillary with precision of 0.79 and mixed neoplasm cancer status (14 tumor free and 13 with tumor) and contains 9 cases of having other malignancy histological type. 242 out of 310 HNSC patients are in C2 and characteristics with low and median histological grade (g1,g2,g3 - 25/26, 144/189, 60/79) mixed neoplasm cancer status(148 tumor free and 71 with tumor) and only 10 cases with other maliganacy histological type specified. 60 other HNSC patients are grouped in another squamous-like cluster C11 that contains 246 out of 340 LUSC patients. LUSC mixed neoplasm cancer status(19 tumor free and 5 with tumor)

C11-LUSC-enriched-squamous-like C11 contains 32 BLCA 26 BRCA, 60 HNSC, 29 LUAD, and 246 LUSC cancer patients. BLCA: mixed papillary diagnosis subtypes (14 papillary 16 non-papillary) BRCA: 24 cases are labeled icd-o-3 histology of 8500/3; progesterone receptor status is mostly negative (20 negative 4 positive); estrogen receptor status is mostly negative (16 negative , 8 positive); 24 histological type is infiltrating ductal carcinoma; HER2 status is more negative (14 negative, 3 positive, 3 equivocal); HNSC: mixed neoplasm cancer status (tumor free: 36 and with tumor 18); contains median histological grade (41 g2 and 15 g3); LUAD: 24 has icd-10 code c34.1 (precision 0.8) and 5 has code c34.3. 23 cases with icd-o-3 histology of 8140/3 (precision 0.77) LUSC: 128 patients have icd-10 code of c34.1 and 90 has code c34.3 and 21 c34.9; tend to be skewed towards patients with higher level of smoking history (level 4 - 130/174, level 3 - 43/67, level 2 61/79 and level 1 - 6/13); better prognosis (tumor free 142/198 and with tumor 38/56);

3.4.3 Survival analysis

We performed survival analysis for each COCA subtype we acquired from Section 3.4.1, using the Cox proportional hazards regression model in the R survival package. We use right-censored survival data for patients: days to death for dead patients, and days to last contact for alive patients as right-censored data. To see how the network constraint affects decomposition result, we impose three different level of network regularization, λ_g: 0 (not constrained), 0.1, and 1.



(a) $\lambda_g = 0$ (b) $\lambda_g = 0.1$ (c) $\lambda_g = 1$

Fig. 9. Predicted survival curves for COCA-clustered patients. x-axis is survival time (day) and y-axis is survival rate.

The log-rank statistics are computed to how well the survival curves are distinguished from each other. the log-rank statistics value when $\lambda_g = 0$ is 409. When $\lambda_g = 0.1$ and 1, the log-rank statistics value are 1151 and 1185, respectively, are higher than that for $\lambda_g = 0$.

3.4.4 Platform factor analysis

We analyze the degree of influence for each platform (Table 1). Norm of a slice of $\mathcal{G} \times_3 \mathbf{U}^{(3)}$ corresponding to each platform is shown in Figure 11. Note that P1(miRNA) is the most dominant, and P4(mRNA) is the next dominant platform for our dataset.

3.4.5 Personalized subtype analysis

For patient i, SNeCT gets patient profile $\mathbf{u}_i^{(1)}$. We calculate the subtype matrix $\mathcal{S} = \mathcal{G} \times_1 \mathbf{u}_i^{(1)}$ ($\in \mathbb{R}^{J_2 \times J_3}$). $\mathcal{S}$ provides a personalized weight information for subtypes for the gene and the platform modes. For a sample patient, Figure 12 shows the heatmap of $\mathcal{S}$, and Figure 14 shows the 3-D visualization of $\mathcal{S}$.

Each row of $\mathcal{S}$ represents a subtype for gene mode, thus norm of each rows represents the influence of each subtype to the patient. Figure 10 shows the calculated subtype weights for 13 representative patients for clusters from COCA (Section 3.4.1).

Each column of $\mathcal{S} \times_3 \mathbf{U}^{(3)}$ represents the platform mode. Norm of each column shows the influence of each platform to the patient i. Figure 13 shows the different influence of platforms for the 13 representative patients for clusters from COCA.

3.5 Prediction

3.5.1 Prediction accuracy of top-k search

SNeCT finds the query profile vector $\mathbf{u}_q$ when query data for a new patient arrives. We select 10% of observed patients in the data as test set. SNeCT conducted factorization excluding the test set. We assume that each patient in the test set arrives time to time as a query, and calculated the top-10 patients for the query patient.

3.5.2 Search speed

SNeCT finds the factors for a query with high speed of parallel SGD. Naive search using the query tensor and raw data takes 85560ms, and SNeCT takes 550ms on average.

4 Conclusion

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Funding

This work has been supported by the... Text Text Text Text.

References

Carlson,A., Betteridge,J., Kisiel,B., Settles,B., Jr.,E.R.H. and Mitchell,T.M. (2010) Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.

Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,Ö., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research,* **39** (SUPPL. 1).

Chang,K., Yih,W. and Meek,C. (2013) Multi-relational latent semantic analysis. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL* pp. 1602–1612.

Choi,D., Jang,J.G. and Kang,U. (2017) Fast, accurate, and scalable method for sparse coupled matrix-tensor factorization. *arXiv preprint arXiv:1708.08640,* **abs/1708.08640**.

Hoadley,K.a., Yau,C., Wolf,D.M., Cherniack,A.D., Tamborero,D., Ng,S., Leiserson,M.D.M., Niu,B., McLellan,M.D., Uzunangelov,V., Zhang,J., Kandoth,C., Akbani,R., Shen,H., Omberg,L., Chu,A., Margolin,A.a., Van'T Veer,L.J., Lopez-Bigas,N., Laird,P.W., Raphael,B.J., Ding,L., Robertson,a.G., Byers,L.a., Mills,G.B., Weinstein,J.N., Van Waes,C., Chen,Z., Collisson,E.a., Benz,C.C., Perou,C.M. and Stuart,J.M. (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell,* **158** (4), 929–944.

Jeon,B., Jeon,I., Sael,L. and Kang,U. (2016*a*) SCouT: Scalable coupled matrix-tensor factorization - algorithm and discoveries. In *32nd IEEE International Conference on Data Engineering (ICDE2016)*, Helsinki, Finland.

Jeon,I., Papalexakis,E.E., Faloutsos,C., Sael,L. and Kang,U. (2016*b*) Mining billion-scale tensors: algorithms and discoveries. *VLDB J.,* **25** (4), 519–544.

Kim,D., Shin,H., Song,Y.S. and Kim,J.H. (2012) Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform,* **45**, 1189–1191.

Kim,S., Sael,L. and Yu,H. (2014) Identifying cancer subtypes based on somatic mutation profile. In *Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics* DTMBIO '14 pp. 19–22 ACM, New York, NY, USA.

Kim,S., Sael,L. and Yu,H. (2015) A mutation profile for top- k patient search exploiting Gene-Ontology and orthogonal non-negative matrix factorization. *Bioinformatics,* **31** (22), 3653–3659.

Kolda,T. and Bader,B. (2006) The TOPHITS model for higher-order web link analysis. *Workshop on Link Analysis, Counterterrorism and Security,* **7**, 26–29.

Kolda,T.G. and Bader,B.W. (2009) Tensor decompositions and applications. *SIAM review,* **51** (3), 455–500.

Kolda,T.G. and Sun,J. (2008) Scalable tensor decompositions for multi-aspect data mining. In *Proceedings - IEEE International Conference on Data Mining, ICDM* pp. 363–372.

Li,W.J. and Yeung,D.Y. (2009) Relation regularized matrix factorization. In *21ST INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI-09), PROCEEDINGS* p. 1126.

Louhimo,R. and Hautaniemi,S. (2011) CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics,* **27**, 887–888.

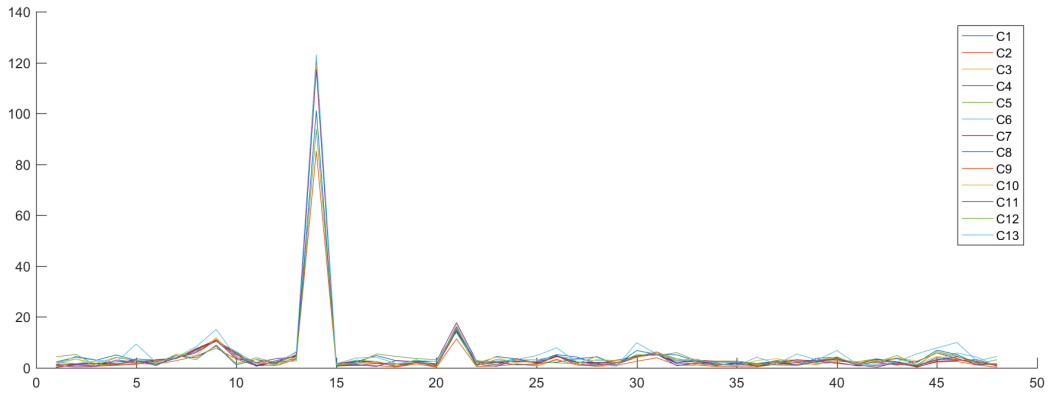


Fig. 10. Norm of gene subtypes.

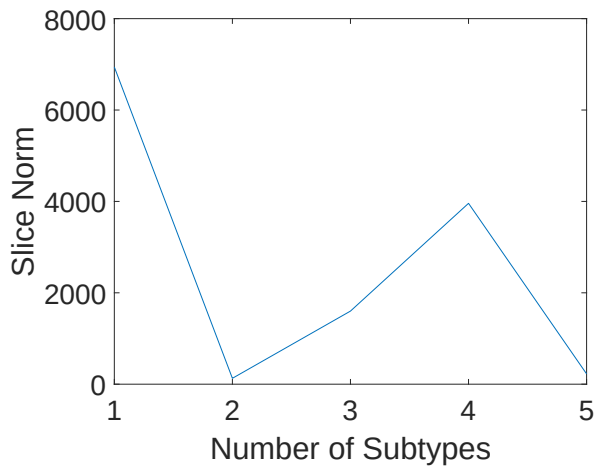


Fig. 11. Norm of slices of $\mathcal{G} \times_3 \mathbf{U}^{(3)}$.

Mankoo,P.K., Shen,R., Schultz,N., Levine,D.A. and Sander,C. (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One,* **6** (11), e24709.

Maruhashi,K., Guo,F. and Faloutsos,C. (2011) MultiAspectForensics: Pattern mining on large-scale heterogeneous networks with tensor analysis. In *Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011* pp. 203–210.

Narita,A., Hayashi,K., Tomioka,R. and Kashima,H. (2012) Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery,* **25** (2), 298–324.

Nickel,M., Tresp,V. and Kriegel,H. (2012) Factorizing YAGO: scalable machine learning for linked data. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012* pp. 271–280.

Omberg,L., Ellrott,K., Yuan,Y., Kandoth,C., Wong,C., Kellen,M.R., Friend,S.H., Stuart,J., Liang,H. and Margolin,A.a. (2013) Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nature Genetics,* **45** (10), 1121–1126.

Recht,B., Re,C., Wright,S. and Niu,F. (2011) Hogwild: a lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems* pp. 693–701.

Sael,L., Jeon,I. and Kang,U. (2015) Scalable tensor mining. *Big Data Research,* **2** (2), 82–86.

Schafer,M., Schwender,H., Merk,S., Haferlach,C., Ickstadt,K. and Dugas,M. (2009) Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics,* **25**, 3228–3235.

Shin,K., Sael,L. and Kang,U. (2017) Fully scalable methods for distributed tensor factorization. *IEEE Transactions on Knowledge and Data Engineering,* **29** (1), 100–113.

Sohn,K.A., Kim,D., Lim,J. and Kim,J.H. (2013) Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC systems biology,* **7**, S9.

Sun,J., Papadimitriou,S. and Yu,P.S. (2006) Window-based tensor analysis on high-dimensional and multi-aspect streams. In *Proceedings - IEEE International Conference on Data Mining, ICDM* pp. 1076–1080.

The Cancer Genome Atlas Network (2017) Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell,* **169** (7), 1327–1341.e23.

Thomas,J. and Sael,L. (2015) Overview of integrative analysis methods for heterogeneous data. In *The 2015 International Conference on Big Data and Smart Computing (BigComp 2015)* number 1 pp. 266–270.

Thomas,J. and Sael,L. (2016) Maximizing information through multiple kernel-based heterogeneous data integration and applications to ovarian cancer. In *6th International Conference on Emerging Databases (EDB)* pp. 97–100 ACM Press.

Tibshirani,R., Walther,G. and Hastie,T. (2001). Estimating the number of clusters in a data set via the gap statistic.

Yuan,Y., Van Allen,E.M., Omberg,L., Wagle,N., Amin-Mansour,A., Sokolov,A., Byers,L.a., Xu,Y., Hess,K.R., Diao,L., Han,L., Huang,X., Lawrence,M.S., Weinstein,J.N., Stuart,J.M., Mills,G.B., Garraway,L.a., Margolin,A.a., Getz,G. and Liang,H. (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature Biotechnology,* **32** (7), 644–652.

Supplementary Information

0.0.1 Gap Statistics Result

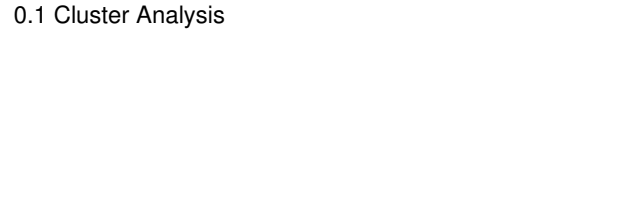


Fig. 1. Gap statistics on patient profiles ($U_(1)$). X-axis is the number of clusters and Y-axis is the gap statistics value. A shows gap statistics for $U_(1)$ obtained with graph regularization value of $\lambda = 1$, B shows that of $U_(1)$ obtained with graph regularization value of $\lambda = 0.1$, and C shows $U_(1)$ obtained without graph regularization.

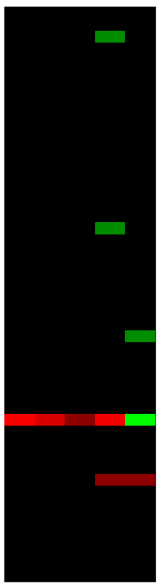


Fig. 12. Heat map of $\mathcal{S}$.

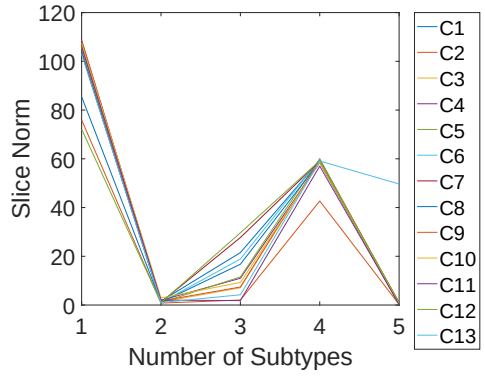


Fig. 13. Personal influence of platforms.

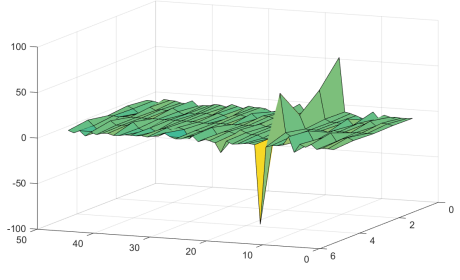


Fig. 14. 3D visualization of $\mathcal{S}$.