Finding Correlated Patterns via High-Order Matching for Multiple Sourced Biological Data

Xi Yang, Guoqiang Han, Jiazhou Chen, and Hongmin Cai

Abstract*—*Objective: **The emergence of multidimensional genomic data poses new challenges in data analysis. Finding correlated patterns within multiple-sourced biological data is useful in understanding potential interactions between the multimodal genomic data. *Methods:* Multidimensional genomic data contain multiple genomic data types, and different types of genomic data have different scales and units. These data cannot simply be aggregated for analysis. To address this issue, a correlated pattern discovery model incorporating prior knowledge is proposed. Tensor similarity is used to measure the correlation between common patterns. The model is combined with prior knowledge, the expression of which is transformed into constraints. Efficient numerical solutions are designed and analyzed. *Results:* The proposed method is shown to perform robustly and effectively with both simulated data and real biological data. We conduct experiments on five real cancer data sets to reveal various cancer subtypes. A survival analysis of these subtypes confirms the effectiveness of the model. *Conclusion:* We introduce a correlated pattern discovery model incorporating prior knowledge. This model is meaningful for the realization of personalized diagnoses by doctors in the treatment of cancer and other diseases. *Significance:* The problem of finding correlated patterns from multiple-sourced biological data was formulated as a high-order graph matching problem, and the prior knowledge data were seamlessly incorporated into the matching model.**

Manuscript received December 18, 2017; revised May 30, 2018; accepted August 12, 2018. Date of publication August 20, 2018; date of current version March 19, 2019. This work was supported in part by the National Natural Science Foundation of China (61472145, 61771007), in part by the Science and Technology Planning Project of Guangdong Province (2016A010101013, 2017B020226004), in part by the Applied Science and Technology Research and Development Project of Guangdong Province (2016B010127003), in part by the Guangdong Natural Science Foundation (2017A030312008), in part by the Fundamental Research Fund for the Central Universities (2017ZD051), and in part by the Health and Medical Collaborative Innovation Project of Guangzhou City (201803010021). *(Corresponding author: Hongmin Cai.)*

X. Yang is with the School of Information Science and Technology, Guangdong University of Foreign Studies, and also with the School of Computer Science and Engineering, South China University of Technology.

G. Han is with the School of Computer Science and Engineering, South China University of Technology.

J. Chen is with the School of Computer Science and Engineering, South China University of Technology, and also with the Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information.

H. Cai is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information, Guangzhou 510006, China (e-mail: hmcai@scut.edu.cn).

Digital Object Identifier 10.1109/TBME.2018.2866266

***Index Terms*—Correlated pattern, matching, multiple-sourced biological data.**

I. Introduction

THE emergence of multi-modal genomic data poses new challenges in data analysis. By integrating various types of genomic data with different scales and units, we can analyse diseases or understand the biological activities from a more comprehensive perspective, thereby reducing the inaccuracy caused by the analysis of single dataset [1]–[4]. Integrated data analysis is also helpful in identifying cancer subtypes and capturing key genomic factors, both of which are pivotal in realizing precision medicine. However, assembling various data types into an informative set for describing complex biological systems remains challenging. The various types of genomic data not only have different units, but also are measured by different measurements. These problems make it difficult to assemble various data types together. The notorious problems of noisy data and missing values further complicate the task.

To simplify the integration of genomic data, researchers have developed various methods, and these are generally categorized into three branches: supervised data integration, semi-supervised data integration, and unsupervised data integration. In supervised models, only labeled samples are used, such as the labels of the genomic data, outcome, or phenotype are known. Various supervised models have been proposed [5]–[8] for biological data integration. In semi-supervised data integration, only a small number of annotated samples and a large number of unlabeled samples are used. such as, graph-based semi-supervised learning is used to predict clinical outcomes in brain cancer and ovarian cancer [9]. Unsupervised integration aims to identify hidden structure within observed data in which neither labels nor phenotypes of interest are known. Sometimes, we want to use unsupervised learning method to integrate data, thus, in this situation, whether the data labled or not, we used as unlabled data. For example, Cao *et al.* [10]–[12] simultaneously analyzed multiple genomic data using canonical correlation analysis. Sparse network-regularized partial least squares (SNPLS) [13] is tailored from the classic theory of canonical correlation analysis and incorporates a gene-gene interaction network to decipher the modular concordance between genes. Yang *et al.* [14] used non-negative matrix factorization to detect patterns in heterogeneous omics multi-modal data. Bayesian-based methods have been presented [15], [16] for modeling multiple genomic data. The iCluster method simultaneously achieves data integration and dimension reduction through a joint latent variable model [17], whereas the Ping-Pong algorithm [18] identifies drug–gene associations. Similarity Network Fusion (SNF) [19] aggregates data types by constructing networks of samples for each available data type, and then efficiently fuses these into one network that represents the full spectrum of underlying data.

The unsupervised methods are applicable to the situation without considering whether the biological data was labled or not,thus, our study is aimed at unsupervised integration. There are still many problems in unsupervised data integration. Popular unsupervised integration models such as iCluster [17] and Similarity Network Fusion [19] are fundamentally dependent on sample-wise similarity. The genomic samples are heavily contaminated by technical noise or contain a large number of missing values. Such imperfections deteriorate the performance of methods based on sample-wise similarities. In addition, early confirmed prior information, such as gene-gene interactions, plays a complementary role in the integrative analysis [20]. Accurately incorporating such prior knowledge would substantially enhance the experimental performances.

The problem of finding correlated patterns is vital in finding disease sub-type from multiple-sourced biological data. In our study, the correlated patterns can be seen as the subtypes of disease. The biological data are typically noisy, and contains many outliers. These characteristic of biological data may reduce the precision of finding correlated patterns by using standard approach. Similar to the model in [21], we formulated the issue as a high order graph matching problem. Our model used "pair-pair" similarity other than popular sample-wise similarity to exploit its robustness over noise contaminations and outliers. The "pair-pair" similarity means the similarity of the paired sample from two datasets. This high-order similarity is quantified by a four-dimensional super-symmetric tensor. Unlike to [21], the prior knowledge on biological data is seamlessly incorporated into the matching model. The prior knowledge is converted into linear constraints of must-link and cannot-link. Rather than using costive power iteration, we demonstrated that the problem could be equivalently solved by a lower-level eigenvalue problem, and we designed numerical algorithm.

To find correlated pattern from multiple-sourced biological data, we adopt high-order similarity measures to quantify the "pair-pair" similarity, which can reduce the impact of noise. A correlated pattern discovery model incorporating prior knowledge is proposed. The proposed model offers four main contributions: 1) The problem of finding correlated patterns from biological data is formulated as a high order graph matching problem; 2) construction of prior functions for combining a priori knowledge with the model; 3) design of numerical algorithm; and 4) presentation of extensive experiments on both simulated and real biological data to demonstrate the robustness and effectiveness of the proposed method.

The remainder of this paper is organized as follows. In Section II, the notation used in this paper and the background to this study are briefly introduced. In Section III, we describe our main tensor matching model for capturing the associations among sets of different types of variables and finding correlated patterns. Section IV presents the results of extensive validations on both simulated and real biological data to demonstrate the superiority of our model over two benchmark models. Finally, the conclusions are given in Section V.

II. Notation and Background

A. Notation

Throughout this paper, we denote vectors, matrices, and tensors in bold lowercase, uppercase, and curlicue uppercase letters, respectively. For a matrix $\boldsymbol{X}$, its i-th row and j-th column are denoted by $\boldsymbol{X}^i$ and $\boldsymbol{X}_j$. We let tensor be denoted by $\boldsymbol{\mathcal{X}}$. A tensor, also known as a multidimensional array, is an element of the tensor product of multiple vector spaces. It is a higher-order generalization of a vector. The N-th order tensor $\mathcal{X} \in R^{I_1, I_2, \ldots, I_N}$ is denoted by $\boldsymbol{\mathcal{X}} = [\mathcal{X}]_{i_1, i_2, \ldots, i_N}, 1 \le i_n \le i_N, 1 \le n \le N$.

The n-mode product of a tensor $\boldsymbol{\mathcal{X}} \in R^{I_1, I_2, \ldots, I_N}$ and a matrix $\boldsymbol{U} \in R^{J_n, I_n}$ is denoted by

$$(\boldsymbol{\mathcal{X}} \times_n \boldsymbol{U})_{i_1, i_2, \ldots, i_{n-1}, j_n, i_n+1, \ldots, i_N} = \sum_{i_n=1}^{I_n} \boldsymbol{x}_{i_1, i_2, \ldots, i_N} \boldsymbol{u}_{j_n, i_n} \tag{1}$$

where the index n indicates that the product applies on the n-th dimension.

III. Methodology

A. Problem Formulation

Popular unsupervised integration models, such as iCluster and SNF are fundamentally dependent on sample-wise similarity. Thus, the subgrouping or clustering is heavily influenced by the metric used to compare samples. However, finding a reliable metric to quantify sample similarity is rather challenging. In particular, sets of genomic samples are heavily contaminated by technical noise and contain a large number of missing values. Such imperfections further deteriorate the performance of methods based on sample-wise similarities.

To tackle with the issue, we model the problem of finding correlated patterns as a graph matching problem. Finding correspondences between visual features (such as interest points, edges, or even raw pixels) is a key problem in many computer vision tasks [21], [22]. The proposed model is inspired by the problem of computer vision matching [21], [22]. For the problem of finding the correlated pattern from multiple sourced biological data, we use the high-order similarity to measure the hyperspace relationship among samples. In particular, by exploiting the similarity between paired samples, the proposed method clusters the genomic data obtained by two different measurements. We model the problem of finding correlated patterns as a tensor matching [21], [22], and wish to cluster the samples into subgroups. The problem can be considered as a matching problem in which samples in the same subgroup possess a high matching value.

Mathematically, let matrices $[\boldsymbol{S}_1]_{m_1 \times n}$ and $[\boldsymbol{S}_2]_{m_2 \times n}$ contain measurements from two different sources for the same n samples. A binary assignment matrix $\boldsymbol{X} = [X]_{i,j}$ is used to measure the match between $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. Moreover, the entry-wise product

TABLE I
THREE FEATURES FOR COMPUTING HIGH-DIMENSIONAL SIMILARITY

Characterization	paired sample $i-j$	Feature
Dot similarity	$\boldsymbol{S}(i)\cdot\boldsymbol{S}(j)$	F^1
Correlation	$\frac{cov(\boldsymbol{S}(i),\boldsymbol{S}(j))}{\sigma(\boldsymbol{S}(i))\sigma_(\boldsymbol{S}(j))}$	F^2
Geometric center	$\boldsymbol{S}(i)-\boldsymbol{S}(j)$	F^3

$\boldsymbol{X}_{i,p}\boldsymbol{X}_{j,q}$ is equal to 1 if and only if the i-th and p-th samples in $\boldsymbol{S}_1$ match to j-th and q-th samples in $\boldsymbol{S}_2$, respectively. In noting this, we attempt to find the optimal "sample-to-sample" match by measuring the higher-order "pair-to-pair" similarity. In particular, the similarity of the paired sample is quantified by a four-dimensional super-symmetric tensor $\boldsymbol{\mathcal{H}}=[\boldsymbol{\mathcal{H}}]_{i,p,j,q}$. Each entry $\boldsymbol{\mathcal{H}}_{i,p,j,q}$ will have a high value if the pair of samples $\{i-j\}$ in $\boldsymbol{S}_1$ is matched to the pair of samples $\{p-q\}$ in $\boldsymbol{S}_2$. This tensor is defined by

$$\boldsymbol{\mathcal{H}}_{i,p,j,q}=\exp(-\gamma\|F_1-F_2\|^2) \tag{2}$$

where $\gamma>0$ is a penalty coefficient. $\boldsymbol{F}_i$ for $i=1,2$ measures the high-order characteristics of the paired samples in $\boldsymbol{S}_i$. They serve as criteria for computing the high-dimensional similarity in addition to the standard one-dimensional "sample-to-sample" similarity. In our experiments, Three features F^1, F^2, F^3 are computed in $\boldsymbol{S}_i$ for representing $\boldsymbol{F}_i$, where $\boldsymbol{F}_1=\{F_1^1,F_1^2,F_1^3\}$ and $\boldsymbol{F}_2=\{F_2^1,F_2^2,F_2^3\}$. The three features used in our experiments are listed in Table I.

In this paper, the problem of finding correlated patterns is formulated as the maximization of the matching between paired samples:

$$\begin{aligned}&\max_{\boldsymbol{X}}\sum_{i,p,j,q}\boldsymbol{\mathcal{H}}_{i,p,j,q}\boldsymbol{X}_{i,p}\boldsymbol{X}_{j,q}\\ &\text{Subject to:}\\ &\|\boldsymbol{X}\|_F\le\sqrt{t}\end{aligned} \tag{3}$$

where the inequality constraint on the Frobenius norm of $\boldsymbol{X}$ allows multiple matching between samples, with t being the maximum allowed matching number.

Equivalently, the tensor energy-based minimization function (3) can be rewritten in matrix form as:

$$\begin{aligned}&\max_{\boldsymbol{X}}\hat{\boldsymbol{X}}^T\hat{\boldsymbol{H}}\hat{\boldsymbol{X}}\\ &\text{Subject to:}\\ &\|\boldsymbol{X}\|_F\le\sqrt{t}\end{aligned} \tag{4}$$

where $\hat{\boldsymbol{X}}$ denotes the vector in $\boldsymbol{R}^{n^2}$ obtained by concatenating the columns of $\boldsymbol{X}$. $\hat{\boldsymbol{H}}$ is the $n^2\times n^2$ symmetric matrix obtained by unfolding the tensor $\boldsymbol{\mathcal{H}}$. The above turns out to be a classical Rayleigh quotient problem. Some heuristic numerical methods have been designed to solve it effectively. For instance, Ngoc *et al.* [23] designed a tensor block coordinate ascent method for hypergraph matching by guaranteeing a monotonic ascent matching score to give a final discrete assignment matrix. Duchenne *et al.* [21] solved a hypergraph problem using the power method, and then projected the derived eigenvalue matrix onto an assignment matrix.

In our article, the problem of finding correlated patterns is formulated as a high order graph matching problem. Thus, the framework is similar to the one proposed in [21]. Our model used "pair-pair" similarity other than popular sample-wise similarity to exploit its robustness over noise contaminations and outliers. Unlike to [21], considering the situation of data with prior knowledge, the prior knowledge on biological data should be incorporated into the matching model.

B. Formulation of Prior Knowledge Into Linear Constraints

In practice, we may have prior knowledge on the closeness of the samples from two different sources. Thus, the problem of finding correlated patterns with prior knowledge is formulated as:

$$\begin{aligned}&\max_{\boldsymbol{X}}\hat{\boldsymbol{X}}^T\hat{\boldsymbol{H}}\hat{\boldsymbol{X}}\\ &\text{Subject to}\\ &\boldsymbol{C}_{ij}\in\{0,1\}\\ &\|\boldsymbol{X}\|_F\le\sqrt{t}\end{aligned} \tag{5}$$

where $\boldsymbol{C}$ is the prior knowledge; $\boldsymbol{C}_{ij}$ is equal to 1 when $\boldsymbol{S}_1(i)$ has relation with $\boldsymbol{S}_1(j)$, and to 0 otherwise.

To guide the maximization while incorporating the prior knowledge, we can model the prior knowledge into linear constraints in the form of must-link and cannot-link relations. Without loss of generality, let us assume that the first sample in $\boldsymbol{S}_1$ is in the same subgroup as the j-th and k-th samples, i.e., $\boldsymbol{X}(1,j)=\boldsymbol{X}(1,k)$. Accordingly, in vector form, we have the equivalent relations:

$$\hat{\boldsymbol{X}}(1+t*n)=\hat{\boldsymbol{X}}(j+t*n)\hat{\boldsymbol{X}}(1+t*n)=\hat{\boldsymbol{X}}(k+t*n)$$

for $1\le t\le n-1$. The must-link constraint $\boldsymbol{d}_1(1)$ for $\boldsymbol{C}(1)$ is

$$\begin{aligned}\boldsymbol{d}_1=(&\underbrace{w,0,\ldots,\overbrace{-1}^{j},0,\ldots,\overbrace{-1}^{k},0,\ldots,}_{n}\\ &\underbrace{w,0,\ldots,\overbrace{-1}^{j+n},0,\ldots,\overbrace{-1}^{k+m1},0,\ldots,}_{n}\\ &\ldots,\\ &\underbrace{w,0,\ldots,\overbrace{-1}^{j+(n-1)*n},0,\ldots,\overbrace{-1}^{j+(n-1)*n},0,\ldots}_{n})\\ &=(d_{11},d_{11},\ldots,d_{11})^T\end{aligned} \tag{6}$$

where w is the degree of the adjacency matrix of $\boldsymbol{C}(1)$ and d_{11} is a vector of size n. In this example, $\boldsymbol{S}_1(1)$ has relations with $\boldsymbol{S}_1(j)$ and $\boldsymbol{S}_1(k)$; thus, the number of relations is 2, $w=2$.

$$\boldsymbol{d}_1{}^T\hat{\boldsymbol{X}}=0 \tag{7}$$

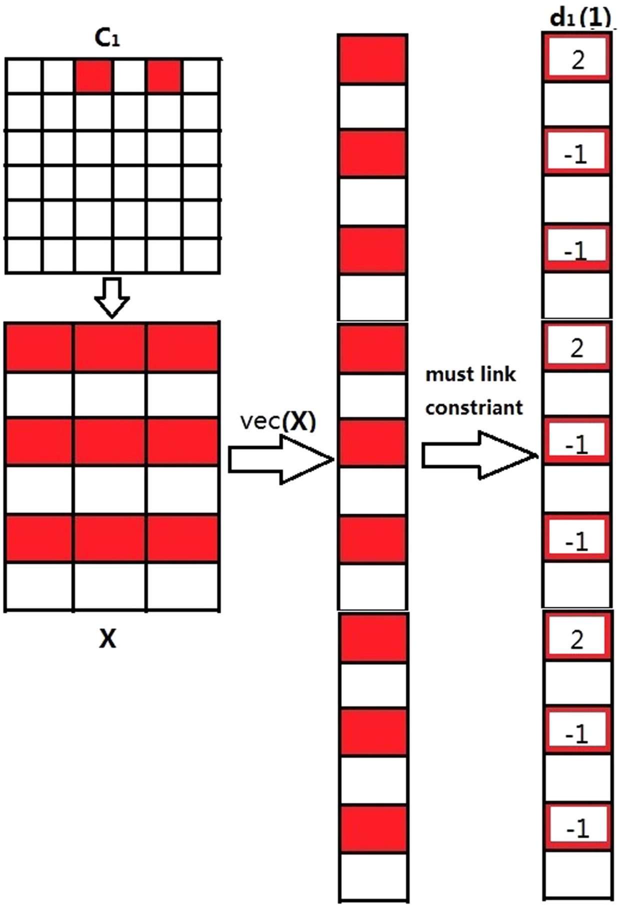

Fig. 1. Illustration of the process of constructing the must-link constraint. We give the example for illustration. First, we get the equivalent relations X of S_1 through the prior knowledge C_1, then we vectorized the equivalent relations X, at last we model vec(X) into linear constraints $d_1(1)$ in the form of must-link and cannot-link relations.

Similarly, we can construct the must-link constraint $\boldsymbol{d}_1(i)$ for $\boldsymbol{S}_1(i)$. $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{m_1}\}$, where $\boldsymbol{d}_i$ is the i-th column of $\boldsymbol{D}$. The process of constructing the must-link constraints is shown in Fig. 1. $\boldsymbol{D}$ can be thought of as the incidence matrix (edges by vertices matrix $\boldsymbol{S}_1$) for the graph of must-link constraints. Thus, we can obtain

$$\boldsymbol{D}^T \hat{\boldsymbol{X}} = 0 \tag{8}$$

Finally, the problem with constraints is generalized to the quadratic problem

$$\max_{\boldsymbol{X}} \hat{\boldsymbol{X}}^T \hat{\boldsymbol{H}} \hat{\boldsymbol{X}}$$

$$\text{Subject to} \tag{9}$$

$$\boldsymbol{D}^T \hat{\boldsymbol{X}} = 0$$

$$\|\boldsymbol{X}\|_F \leq \sqrt{t}$$

The workflow of the proposed model is shown in Fig. 2. The two input matrices shown in Fig. 2(a) represent the two different data types of the same samples; for ease of illustration, the sample number is 3. The similarity measure H computed from the pair of the first matrix and the pair of the second matrix is shown in Fig. 2(b). Samples i and j from the first matrix are correlated with samples i' and j' from the second matrix. The assignment matrix X is shown in Fig. 2(c). A correlated pattern can be found by matching.

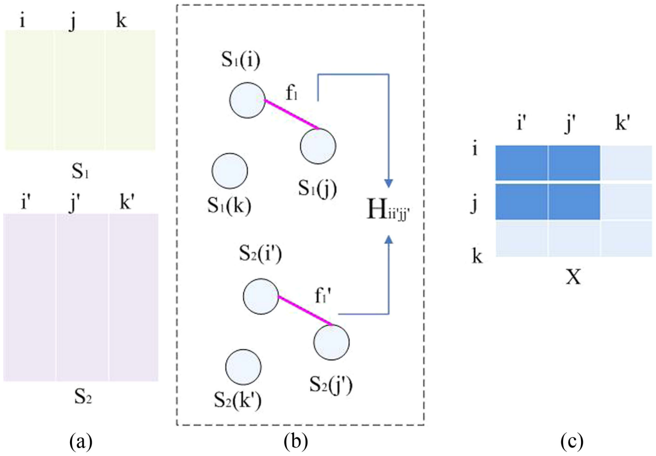

Fig. 2. Illustration of the workflow of the proposed model. The correlated pattern of the two matrices S_1, S_2 is obtained by the proposed high-order matching model. (a) The two input matrices. These matrices represent the two different data types of the same samples; for ease of illustration, the sample number is 3. (b) The similarity measure H computed from the pair of the first matrix and the pair of the second matrix. Samples i and j from the first matrix are correlated with samples i' and j' from the second matrix. (c) The assignment matrix X. A correlated pattern is found by matching.

C. Numerical Solution

The maximization problem (9) is a quadratic convex optimization that can be solved by the gradient descent method. However, the pair similarity tensor $\boldsymbol{\mathcal{H}}$ has order $n \times n \times n \times n$, with n being the sample size. This introduces a large computation and storage burden. It is surprising to note that the introduction of the linear constraints will effectively alleviate this problem, as shown below.

The Lagrange equation for (9) is given by

$$L(\boldsymbol{X}, \lambda, \boldsymbol{\mu}) = \hat{\boldsymbol{X}}^T \hat{\boldsymbol{H}} \hat{\boldsymbol{X}} - \lambda(\hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}} - t) + 2\boldsymbol{\mu}^T \boldsymbol{D}^T \hat{\boldsymbol{X}} \tag{10}$$

where $\lambda, \boldsymbol{\mu}$ are Lagrange multipliers.

Differentiating the Lagrange equation with respect to $\hat{\boldsymbol{X}}$, we obtain

$$\hat{\boldsymbol{X}}^T \hat{\boldsymbol{H}} - \lambda \hat{\boldsymbol{X}}^T + \boldsymbol{\mu}^T \boldsymbol{D}^T = 0 \tag{11}$$

Multiplying the transpose of (11) on the left by $\boldsymbol{D}^T$ and using the condition that $\boldsymbol{D}^T \boldsymbol{X} = 0$, we have

$$\boldsymbol{D}^T \boldsymbol{H}^T \boldsymbol{X} + \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{\mu} = 0$$

Therefore, at the optimal solution $\hat{\boldsymbol{X}}^*$, one has

$$\boldsymbol{\mu} = -(\boldsymbol{D}^T \boldsymbol{D})^{-} \boldsymbol{D}^T \hat{\boldsymbol{H}}^T \hat{\boldsymbol{X}}^* \tag{12}$$

with $\boldsymbol{D}^-$ being the pseudo-inverse of $\boldsymbol{D}$. Substituting (12) into (11), we then obtain

$$\boldsymbol{P}\hat{\boldsymbol{H}}^T\hat{\boldsymbol{X}}^* = \lambda\hat{\boldsymbol{X}}^* \tag{13}$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{D})^-\boldsymbol{D}^T = \boldsymbol{I} - \boldsymbol{D}\boldsymbol{D}^-$. Therefore, the optimal solution $\hat{\boldsymbol{X}}^*$ is one of the eigenvectors of the matrix $\boldsymbol{P}\hat{\boldsymbol{H}}^T$.

Note that $\boldsymbol{P}^2 = \boldsymbol{P}$, so that $\boldsymbol{P}$ is a projection matrix. As the matrix $\boldsymbol{P}\hat{\boldsymbol{H}}^T$ and the symmetric matrix $\boldsymbol{P}\hat{\boldsymbol{H}}\boldsymbol{P}$ share the same eigenvectors, we have

$$eig(\boldsymbol{P}\boldsymbol{H}^T) = eig(\boldsymbol{P}^2\hat{\boldsymbol{H}}^T) = eig(\boldsymbol{P}\hat{\boldsymbol{H}}^T\boldsymbol{P}) \tag{14}$$

Instead of obtaining the eigenvectors from the matrix $\boldsymbol{P}\hat{\boldsymbol{H}}^T$, we wish to find them via the symmetric matrix $\boldsymbol{P}\hat{\boldsymbol{H}}^T\boldsymbol{P}$.

Assume that the orthogonal decomposition of $\boldsymbol{D}$ is

$$\boldsymbol{D} = \boldsymbol{Q}\begin{bmatrix} \boldsymbol{R} & \boldsymbol{S} \\ 0 & 0 \end{bmatrix} \tag{15}$$

where $\boldsymbol{R}$ is an upper-triangular matrix of order r, $\boldsymbol{S}$ is an $r \times (p - r)$-order matrix, and $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_n$. Thus,

$$eig(\boldsymbol{P}\hat{\boldsymbol{H}}^T\boldsymbol{P}) = eig(\boldsymbol{J}_{n-r}\boldsymbol{Q}^T\hat{\boldsymbol{H}}\boldsymbol{Q}\boldsymbol{J}_{n-r}) \tag{16}$$

with

$$\boldsymbol{J}_{n-r} = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{I}_{n-r} \end{bmatrix}$$

Suppose that $\boldsymbol{Q}^T\hat{\boldsymbol{H}}\boldsymbol{Q} = [\begin{smallmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{smallmatrix}]$. Then, we can verify that $\boldsymbol{J}\boldsymbol{Q}^T\hat{\boldsymbol{H}}\boldsymbol{Q}\boldsymbol{J} = [\begin{smallmatrix} 0 & 0 \\ 0 & \boldsymbol{G}_{22} \end{smallmatrix}]$.

Finally, we have

$$eig(\boldsymbol{P}\hat{\boldsymbol{H}}^T\boldsymbol{P}) = eig(G_{22}) \tag{17}$$

Hence, the optimal solution $\hat{\boldsymbol{X}}^*$ for the constrained eigenvalue problem (4) is simply the eigenvector of the matrix $\boldsymbol{G}_{22}$. In particular, $\boldsymbol{z}$ is the eigenvector corresponding to the minimum eigenvalue, i.e.,

$$\boldsymbol{G}_{22}\boldsymbol{z} = \lambda_{min}\boldsymbol{z} \tag{18}$$

The solution to (4) is

$$\hat{\boldsymbol{X}}^* = \begin{bmatrix} \boldsymbol{G}_{12} \\ \boldsymbol{G}_{22} \end{bmatrix}\boldsymbol{z} \tag{19}$$

Rather than using costive power iteration, we demonstrated that the problem could be equivalently solved by a lower-level eigenvalue problem. We sketch the pseudocode in Algorithm 1.

Algorithm 1: Algorithm for Solving the Proposed Model (5).

```
Input: $S_1$, $S_2$
  Find the pair of matrices $S_1$ and $S_2$.
  Compute $F_i$ by the criteria for computing
    high-dimensional similarity.
  Compute the similarity measure $H$ by solving (2).
    Find the nearest $k$ pairs of matrix $S_2$ that are similar
    to each pair of matrix $S_1$ according to the similarity
    measure.
  Construct the must-link constraint of prior knowledge
    by solving (8).
  Compute the assignment matrix $X$ by solving (19).
  Correlated matching.
Output: The correlated pattern.
```

IV. Results and Discussion

To demonstrate the performance of the proposed method, extensive experiments on both simulated and real datasets were conducted. The two well-known methods of High-Order Graph Matching (GM) [21] and SNF [19] were used as benchmark models to compare the performance of our method.

GM [21] uses higher-order constraints instead of unary or pairwise ones. This method generalizes the spectral matching method to higher-order potentials: the corresponding hypergraph matching problem is formulated as the maximization of a multilinear objective function over all permutations of the features. This function is defined by a tensor representing the affinity between feature tuples. It is maximized using a multidimensional power method to solve a relaxed version of the problem, and this solution is then projected onto the closest assignment matrix.

SNF [19] aggregates data types by constructing networks of samples (e.g., patients) for each available data type, and then efficiently fuses these into one network that represents the full spectrum of underlying data. For example, to create a comprehensive view of a disease given a cohort of patients, SNF computes and fuses patient similarity networks obtained from each separate data type, taking advantage of the complementarity of the data. In SNF, there are two free parameters, η and K. SNF is not sensitive to these two parameters [19]. A reasonable range for η would be 0.3–1. The rule of thumb for K is $K = N/C$, where N is the number of patients and C is the number of clusters thought to be in the data. However, if C is unknown, we set $K \approx N/10$.

For quantitative performance comparisons, the accuracy and Normalized Mutual Information (NMI) [19] metrics were used. The accuracy is defined as:

$$Accuracy = \frac{m_l}{n}$$

where n is the sample size and m_l is the number of accurately matched samples categorized into the same cluster. For the reason of SNF did not to do matching, the accuracy is only used to measure the performance of our method and GM method.

For comparing the proposed method with the two methods, GM method and SNF method, we brought NMI to measure. The NMI for two clustering results U and V is defined as

$$NMI = \frac{I(U, V)}{H(U)H(V)}$$

where $I(U, V)$ is the mutual information between U and V, and $H(\cdot)$ is the entropy. NMI takes values from 0–1. It measures the concordance between two clustering results. A higher NMI value implies better agreement between the two clustered results.

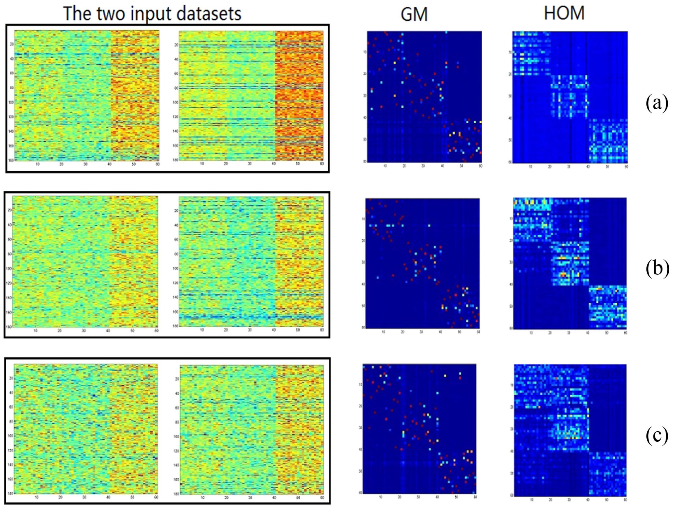

Fig. 3. Illustrative results of the simulation experiment with regard to the three correlated patterns. Each row presents results obtained in different noise conditions. From top to bottom, the two input matrices were contaminated by low-level noise (signal-to-noise ratio (SNR) ≈ 7), intermediate-level noise (SNR ≈ 3), and high-level noise (SNR ≈ 0.6). (a) Results for GM and HOM for low-level noise (SNR ≈ 7). (b) Results for GM and HOM for intermediate-level noise (SNR ≈ 3). (c) Results for GM and HOM for high-level noise (SNR ≈ 0.6). The proposed method clearly shows the three diagonal patterns when the noise level is not high (SNR ≈ 7 and SNR ≈ 3). The results for the high noise level also clearly exhibit the diagonal pattern.

TABLE II
ACCURACY OBTAINED BY HOM AND GM ON SIMULATED DATASET I WITH DIFFERENT PERCENTAGES OF IRRELEVANT FEATURES AND VARIOUS NOISE CONDITIONS

	irrelevant proportion			noise level		
method	0.1	0.2	0.3	snr=7	snr=3	snr=0.6
GM	0.9833	0.9667	0.8833	0.9833	0.9500	0.8333
HOM	**1.0000**	**0.9667**	**0.9167**	**1.0000**	**1.0000**	**0.8833**

TABLE III
NMI OBTAINED BY THREE METHODS ON SIMULATED DATASET I WITH DIFFERENT PERCENTAGES OF IRRELEVANT FEATURES AND VARIOUS NOISE CONDITIONS

	irrelevant proportion			noise level		
method	0.1	0.2	0.3	snr=7	snr=3	snr=0.6
GM	0.7118	0.8411	0.7118	0.7118	0.7317	0.4751
SNF	1.0000	1.0000	1.0000	1.0000	**1.0000**	0.7036
HOM	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.8411	**0.8795**

A. Comparative Studies on Simulated Data

We designed three simulated experiments to test the performance of the proposed method. Data were created to simulate two datasets from heterogeneous sources. To compare the robustness of the proposed method to different numbers of correlated patterns, the simulated datasets in the first two experiments have different number of common patterns. The two datasets in the first experiment had three block structures representing three common patterns. The two datasets in the second experiment had five block structures representing five correlated patterns. To consider the noise-contaminated characteristic, the simulated datasets were contaminated by different level of Gaussian white noise in each experiment. To test the robust of the proposed method to the impact of irrelevant samples in each experiment, different percent of irrelevant biological samples were simulated in datasets. At last, the effectiveness of the model with the prior knowledge is illustrated in the third experiment.

1) Results of the Simulation Experiment With Regard to the Three Correlated Patterns: The two input matrices in the first experiment were $(\boldsymbol{S}_1)_{180\times 60} = [\boldsymbol{S}_1(1), \boldsymbol{S}_1(2), \ldots, \boldsymbol{S}_1(60)]$ and $(\boldsymbol{S}_2)_{180\times 60} = [\boldsymbol{S}_2(1), \boldsymbol{S}_2(2), \ldots, \boldsymbol{S}_2(60)]$. Each dataset had three block structures representing three correlated patterns. The proposed model (with different percentages of prior knowledge) and the benchmark models were employed to analyze the simulated data. To compare the robustness of the proposed method in the case of noise contamination, the values of the input matrices were adjusted to simulate data degradation by various levels of noise. The results of the simulation experiment with regard to the various levels of noise are shown in Fig. 3. These heatmaps enable a visual comparison. The data construction described above meant that the three diagonal patterns in matrix $\boldsymbol{X}$ possessed a high matching value, whereas other regions were subject to noise perturbation with smaller matching values. From the top to the bottom, the input matrices were contaminated by three different noise levels. The second and third columns show the matching results identified by GM and high-order matching (HOM), respectively.

For the first experiment with a correlated pattern of size three, the assignment matrix X given by the proposed model clearly exhibits the three diagonal patterns of the matrix (Fig. 3(a) and Fig. 3(b)). When the data were highly contaminated by noise, the assignment matrix X produced by the proposed model shows the three diagonal patterns of the matrix with minor perturbations (Fig. 3(c)). The comparative numerical results about matching accuracy are summarized in Tables II, and the comparative numerical results about NMI are summarized in Tables III. The proposed method achieved superior performance, as shown by the higher accuracy compared with GM in every case and the higher NMI compared with GM and SNF in almost every case.

2) Results of the Simulation Experiment With Regard to the Five Correlated Patterns: To compare the robustness of the proposed method to different numbers of correlated patterns, a second experiment was conducted. The two input matrices in the second experiment were $(\boldsymbol{S}_1)_{260\times 200} = [\boldsymbol{S}_1(1), \boldsymbol{S}_1(2), \ldots, \boldsymbol{S}_1(200)]$ and $(\boldsymbol{S}_2)_{260\times 200} = [\boldsymbol{S}_2(1), \boldsymbol{S}_2(2), \ldots, \boldsymbol{S}_2(200)]$. These two matrices were constructed with five correlated patterns. The proposed model (with different percentages of prior knowledge) and the two benchmark models were employed to analyze the test data. To compare the robustness of the proposed method to noise contamination, the input matrices were adjusted to simulate the degradation of data under various levels of noise. The results of the simulation experiment with regard to the various levels of noise are shown in Fig. 4. Again, these heatmaps enable a visual comparison. The data construction described above resulted in the five diagonal patterns of the matrix $\boldsymbol{X}$ possessing a high matching value, whereas other regions were subject to noise perturbation and had smaller

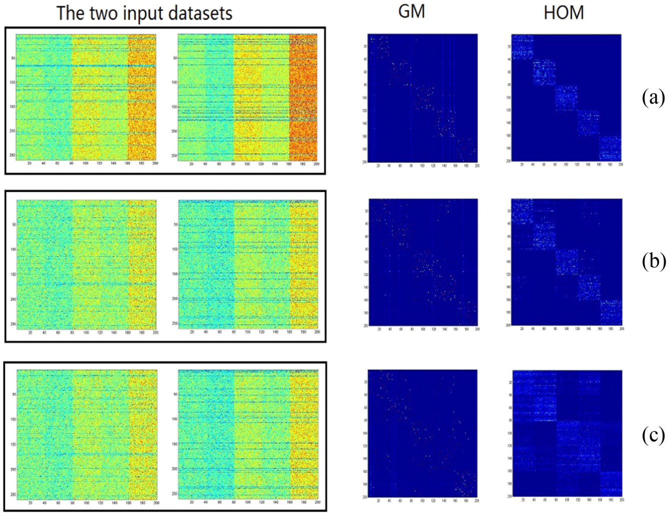

Fig. 4. Illustrative results of the simulation experiment with regard to the five correlated pattern. Each row presents results obtained in different noise conditions. From top to bottom, the two input matrices were contaminated by low-level noise (SNR ≈ 7), intermediate-level noise (SNR ≈ 3), high-level noise (SNR ≈ 0.6). (a) Results for GM and HOM for low-level noise (SNR ≈ 7). (b) Results for GM and HOM for intermediate-level noise (SNR ≈ 3). (c) Results for GM and HOM for high-level noise (SNR ≈ 0.6). The proposed method clearly shows the five diagonal patterns when the noise level is not high (SNR ≈ 7 and SNR ≈ 3). The results for the high noise level also clearly exhibit the diagonal pattern.

TABLE IV
ACCURACY OBTAINED BY HOM AND GM ON SIMULATED DATASET II WITH DIFFERENT PERCENTAGES OF IRRELEVANT FEATURES AND VARIOUS NOISE CONDITIONS

	irrelevant proportion			noise level		
method	0.1	0.2	0.3	snr=7	snr=3	snr=0.6
GM	1.0000	0.9950	**1.0000**	0.9850	0.9200	0.6600
HOM	**1.0000**	**1.0000**	0.9950	**0.9950**	**0.9450**	**0.7350**

TABLE V
NMI OBTAINED BY THREE METHODS ON SIMULATED DATASET II WITH DIFFERENT PERCENTAGES OF IRRELEVANT FEATURES AND VARIOUS NOISE CONDITIONS

	irrelevant proportion			noise level		
method	0.1	0.2	0.3	snr=7	snr=3	snr=0.6
GM	0.9605	0.8299	0.8238	0.9605	0.6832	0.4751
SNF	0.8096	0.8096	0.8096	0.8096	0.7916	**0.7726**
HOM	**1.0000**	**0.9709**	**0.9709**	**1.0000**	**0.8659**	0.7677

matching values. The second and the third columns show the matching results identified by GM and HOM, respectively.

For the second experiment with a correlated pattern of size five, the assignment matrix X given by the proposed model clearly shows five diagonal patterns (Fig. 4(a) and Fig. 4(b)). When the data were highly contaminated by noise, the assignment matrix X given by the proposed model exhibits the five diagonal patterns of the matrix with minor perturbations (Fig. 4(c)). The comparative numerical results about matching accuracy are summarized in Table IV, and the comparative numerical results about NMI are summarized in Table V. The proposed method achieved superior performance, as shown by the higher accuracy compared with the GM method in every case and the higher NMI compared with GM and SNF in almost every case.

TABLE VI
ACCURACY OBTAINED BY HOM ON SIMULATED DATASETS I AND II WITH DIFFERENT PERCENTAGES OF PRIOR KNOWLEDGE

Prior knowledge	simulated data I	simulated data II
0.25	**0.8500**	0.7350
0.50	**0.8833**	0.7450
0.75	**0.9333**	0.7450

TABLE VII
SUMMARY STATISTICS OF THE FIVE CANCER DATASETS

Cancer Type	Patients	mRNA Expression	DNA Methylation
GBM	**215**	12042	1491
BIC	**105**	17814	23094
KRCCC	**122**	17899	24960
LSCC	**106**	12042	23074
COAD	**92**	17814	23088

3) The Effectiveness of the Model With the Prior Knowledge: To illustrate the influence of the prior knowledge, the comparative numerical matching results for datasets I and II with different percentages of prior knowledge are summarized in Table VI. We construct the different must-link constraints according to the different prior knowledge in the proposed model. From these results, we can see that higher levels of prior knowledge lead to higher matching accuracy. This experiment shows that adding prior knowledge to the model is meaningful in terms of finding association patterns, and demonstrates that a priori knowledge improves the accuracy of data matching.

B. Comparative Studies on Real Biological Data

We applied the proposed method to five cancer profiles from The Cancer Genome Atlas [19]: glioblastoma multiforme (GBM), breast invasive carcinoma (BIC), kidney renal clear cell carcinoma (KRCCC), lung squamous cell carcinoma (LSCC), and colon adenocarcinoma (COAD). The DNA methylation and mRNA expression data for these cancers vary in terms of sample size (from 92 for COAD to 215 for GBM) and number of measurements (from 12042 mRNA in GBM to 23088 methylated genes in COAD), as well as heterogeneity. The number of patients used in our experiments and the number of genes in the mRNA expression and DNA methylation datasets are listed in Table VII.

We found highly correlated patterns corresponding to the clusters in DNA methylation and mRNA expression. The discovery of subtypes is still in the research stage which is lack of standard. For example, one analysis had identified two subtypes from GBM dataset [24], but anther analysis had identified four subtypes from GBM dataset, which does not agree with the previous findings [25]. A recent two approaches [19], [26] had identified three subtypes, however the three subtypes are not the

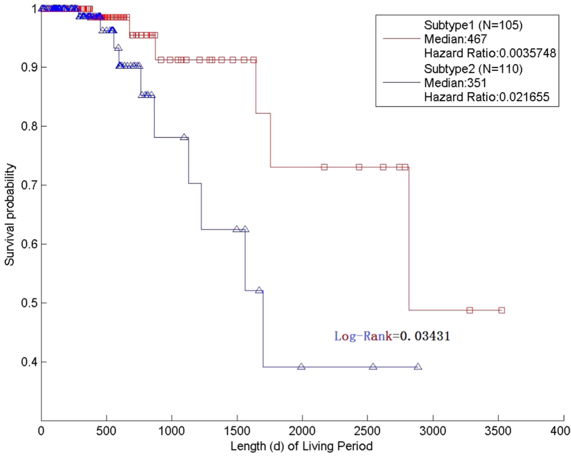

Fig. 5. Survival curves for two subtypes found in GBM.

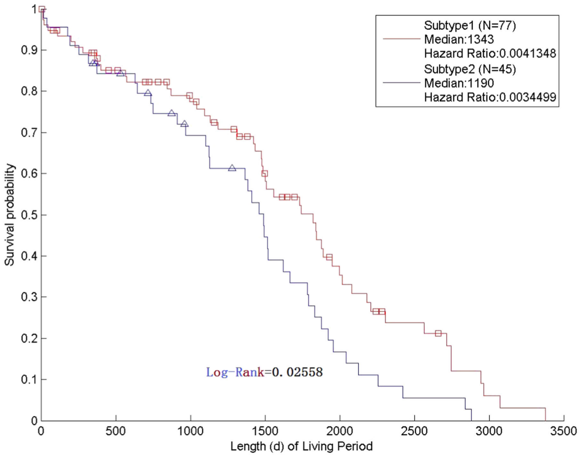

Fig. 7. Survival curves for two subtypes found in KRCCC.

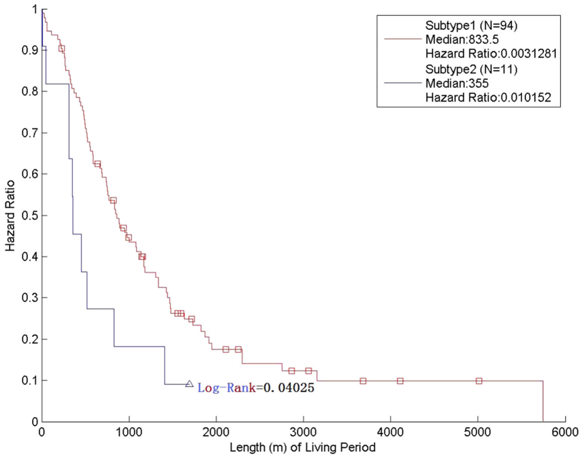

Fig. 6. Survival curves for two subtypes found in BIC.

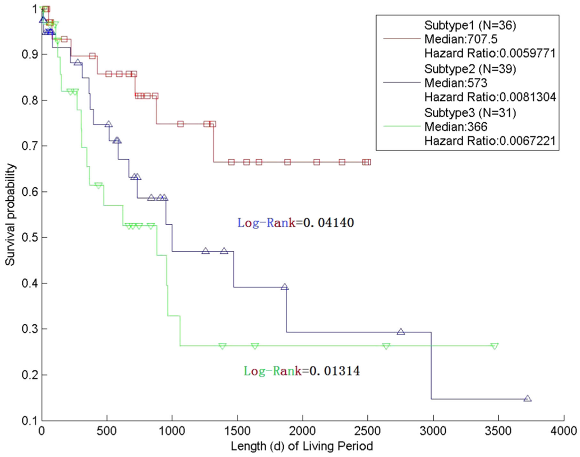

Fig. 8. Survival curves for three subtypes found in LSCC.

same. We found two subtypes in the GBM data, two subtypes in BIC, two subtypes in KRCCC, three subtypes in LSCC, and three subtypes in COAD. We used the Kaplan–Meier method to draw survival curves for the subtypes found in each dataset, and the Log-rank test was used for survival analysis to demonstrate the validity of our model on the five datasets. In the experiment, a P-value of 0.05 was set as a threshold for measuring the significance of subtype survival differences. That is, $P < 0.05$ denotes a significant difference, with smaller P-values indicating more significant differences in survival between subtypes.

The survival curves for the two subtypes found in GBM are shown in Fig. 5. In this figure, subtype 1 and subtype 2 can be clearly distinguished. The P-value is 0.03431, indicating that there are significant differences between these two subtypes. The survival curves for the two subtypes found in BIC are shown in Fig. 6. In this case, subtype 1 and subtype 2 can be clearly distinguished, and the P-value is 0.04025. This implies that there are significant differences between the two subtypes. The survival curves for the two subtypes found in KRCCC are shown in Fig. 7. From this figure, it can be seen that subtype 1 and subtype 2 are clearly distinguishable. The P-value is 0.02558, indicating that there are significant differences between the two subtypes. The survival curves for the three subtypes found in LSCC are shown in Fig. 8. For subtype 1 and subtype 2, the P-value is 0.04140, and that for subtype 1 and subtype 3 is 0.01314, for subtype 2 and subtype 3 is 0.53. Thus, there is a significant difference between subtypes 1 and 2 and between subtypes 1 and 3. From Fig. 8, we can see that subtype 1 and subtype 2 can be clearly distinguished, and subtype 1 and subtype 3 can be clearly distinguished. But, for subtype 2 and subtype 3, the two subtypes can not be clearly distinguished. The survival curves for the three subtypes found in COAD are shown in Fig. 9. For subtype 1 and subtype 2, the P-value is 0.01536, and that for subtype 2 and subtype 3 is 0.01939. Thus, there is a significant difference between subtypes 1 and 2 and between subtypes 2 and 3. For subtype 1 and subtype 3, the P-value is 0.00974. Again, there is a significant difference between these subtypes. From Fig. 9, we find that the three subtypes can be clearly distinguished.

From the experiments on five biological datasets, we have found that our method can effectively identify differentiated subtypes, which may allow patients to receive personalized treatment strategies.

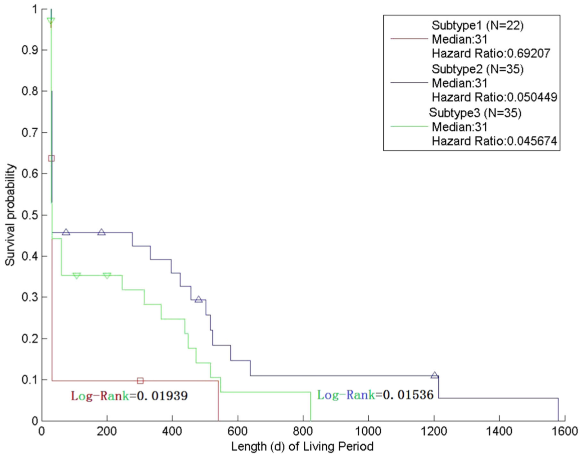

Fig. 9. Survival curves for three subtypes found in COAD.

V. Conclusions

In this paper, we have proposed a correlated pattern discovery model that incorporates prior knowledge. The problem of finding correlated patterns from multiple-sourced biological data was formulated as a matching problem, and tensor similarity was used to measure the correlation among correlated patterns. The proposed model can also be combined with prior knowledge, the expression of which is transformed into constraints. Efficient numerical solutions have been designed and analyzed, and we tested the proposed method extensively with simulated and real biological datasets. The experimental results demonstrate the superior performance of the proposed method in accurately and robustly finding the correlated patterns. This method should allow doctors to realize personalized diagnoses and treatments for cancer and other diseases.

References

[1] U. D. Akavia *et al.*, “An integrated approach to uncover drivers of cancer,” *Cell*, vol. 143, no. 6, pp. 1005–1017, 2011.

[2] J. Menche *et al.*, “Uncovering disease-disease relationships through the incomplete interactome,” *Science*, vol. 347, no. 6224, pp. 841–841, 2015.

[3] C. Kandoth *et al.*, “Integrated genomic characterization of endometrial carcinoma,” *Nature*, vol. 497, no. 7447, pp. 67–73, 2013.

[4] S. Zhang *et al.*, “A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules,” *Bioinformatics*, vol. 27, no. 13, pp. i401–i409, 2011.

[5] M. D. Ritchie *et al.*, “Methods of integrating data to uncover genotype-phenotype interactions,” *Nature Rev. Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[6] P. K. Mankoo *et al.*, “Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles,” *Plos One*, vol. 6, no. 11, 2011, Art. no. e24709.

[7] G. R. Lanck *et al.*, “A statistical framework for genomic data fusion,” *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[8] H. B. Shen *et al.*, “Ensemble classifier for protein fold pattern recognition,” *Bioinformatics*, vol. 22, no. 14, pp. 1717–1722, 2006.

[9] D. Kim *et al.*, “Synergistic effect of different levels of genomic data for cancer clinical outcome prediction,” *J. Biomed. Informat.*, vol. 45, no. 6, pp. 1191–1198, 2012.

[10] D. Cao *et al.*, “Sparse canonical correlation analysis applied to omics studies for integrative analysis and biomarker discovery,” *J. Chemometrics*, vol. 29, no. 6, pp. 371–378, 2015.

[11] D. Lin *et al.*, “Correspondence between FMRI and SNP data by group sparse canonical correlation analysis,” *Med. Image Anal.*, vol. 18, no. 6, pp. 891–902, 2014.

[12] P. Chalise *et al.*, “Simultaneous analysis of multiple data types in pharmacogenomic studies using weighted sparse canonical correlation analysis,” *Omics A J. Integrative Biol.*, vol. 16, nos. 7/8, pp. 363–373, 2012.

[13] J. Chen and S. Zhang, “Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data,” *Bioinformatics*, vol. 32, no. 11, pp. 1724–1732, 2016.

[14] Z. Yang *et al.*, “A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data,” *Bioinformatics*, vol. 32, no. 1, pp. 325–342, 2016.

[15] P. Kirk *et al.*, “Bayesian correlated clustering to integrate multiple datasets,” *Bioinformatics*, vol. 28, no. 24, pp. 3290–3297, 2012.

[16] M. Wu *et al.*, “Simultaneous inference of phenotype-associated genes and relevant tissues from GWAs data via Bayesian integration of multiple tissue-specific gene networks,” *J. Molecular Cell Biol.*, vol. 9, no. 6, pp. 436–452, 2017.

[17] R. Shen *et al.*, “Integrative subtype discovery in glioblastoma using icluster,” *Plos One*, vol. 7, no. 4, 2012, Art. no. e35236.

[18] Z. Kutalik *et al.*, “A modular approach for integrative analysis of large-scale gene-expression and drug-response data,” *Nature Biotechnol.*, vol. 26, no. 5, pp. 531–539, 2008.

[19] B. Wang *et al.*, “Similarity network fusion for aggregating data types on a genomic scale,” *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.

[20] Y. Luo *et al.*, “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 573.

[21] O. Duchenne *et al.*, “A tensor-based algorithm for high-order graph matching,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2383–2395, Dec. 2011.

[22] Q. Zhang *et al.*, “Learning graph matching: Oriented to category modeling from cluttered scenes,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014, pp. 1329–1336.

[23] Q. N. Ngoc *et al.*, “A flexible tensor block coordinate ascent scheme for hypergraph matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5270–5278.

[24] J. M. Nigro *et al.*, “Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma,” *Cancer Res.*, vol. 65, no. 5, pp. 1678–1686, 2005.

[25] G. W. Verhaak *et al.*, “An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR AND NF1,” *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[26] D. Sturm *et al.*, “Hotspot mutations in H3F3A AND IDH1 Define distinct epigenetic and biological subgroups of glioblastoma,” *Cancer Cell*, vol. 22, no. 4, pp. 425–437, 2012.