IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 17, 2019, accepted August 25, 2019, date of publication September 4, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939405

An Integrated Graph Regularized Non-Negative Matrix Factorization Model for Gene Co-Expression Network Analysis

YING-LIAN GAO[1], MI-XIAO HOU[2], JIN-XING LIU[2], (Member, IEEE), AND XIANG-ZHEN KONG[2]

[1]Qufu Normal University Library, Qufu Normal University, Rizhao 276826, China
[2]School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China

Corresponding authors: Jin-Xing Liu (sdcavell@126.com) and Xiang-Zhen Kong (kongxzhen@163.com)

This work was supported in part by the National Science Foundation of China under Grant 61872220, Grant 61572284, and Grant 61702299.

ABSTRACT Studies of cancers have become diversified in recent years, especially with the availability of multi-omics data. Establishing an effective integrative model to process more types of data has become a new research hotspot. In order to conduct deeper mining in cancer, building a gene co-expression network based on multi-omics data for more valuable clues is the useful means in this research. Based on the data-driven problems of cancer networks, this paper proposes an integrated Graph Regularized Non-negative Matrix Factorization model that can be used for network analysis called iGMFNA. We apply iGMFNA to two cancer datasets from The Cancer Genome Atlas (TCGA) for analysis. We demonstrate that our method is indeed more effective than other integrated methods. In terms of network analysis and mining, we also define a multi-measure for nodes in the network to identify cancer-related genes. Through text mining, we verify some genes discovered by iGMFNA.

INDEX TERMS Integrative model, network mining, gene co-expression network, TGCA.

I. INTRODUCTION

With the rapid development of high-throughput technologies, we can accurately obtain various biological sequencing data from organisms at various stages of development. These data, collectively known as multi-omics data, include profiles of gene expression, gene regulation, protein/RNA interactions, mutation, methylation, and so on. Multi-omics data provide the foundation of analysis for the diagnosis and treatment of cancers [1]. In particular, the popular database, The Cancer Genome Atlas (TCGA), aggregates these types of data and provides a powerful channel for the studies of cancers based on network models in recent years [2]–[4].

Many studies have demonstrated the possibility for deeper data mining of multi-omics data. Multi-omics data are measured and collected in different ways; their distributions and noise are variable. Most importantly, they represent different aspects of a biological system, providing a variety of useful views of the complete system. A joint analysis of the same set of sample data from a multi-omics group may get more perceptual results than analyzing a single data and provide a more comprehensive global view of the biological system. However, the variety in multi-omics data also presents a new problem: how to coordinate the differences among the types of data to better reflect the characteristics of cancers. Individual models cannot address these data uniformly and achieve good results. For the analysis of different types of data on the same cancer, many experts and scholars have proposed various integrated models. For example, Shen et al. proposed iCluster to cluster various types of data, which is intended to find new subtypes [5]. Zhang et al. introduced joint Non-negative Matrix Factorization (jNMF) which integrated the basic Non-negative Matrix Factorization (NMF) into an integrated model and identified feature genes [6]. Wang et al. proposed a Similarity Network Fusion (SNF) model to conduct sample fusion network construction for multi-omics data and subtype analysis [3]. Yang et al. employed an integrative NMF (iNMF) approach, which applied the L_1-norm, to detect relevant cancer modules [7]. Stražar et al. introduced an integrative Orthogonality-regularized NMF (iONMF) model,

The associate editor coordinating the review of this article and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam.

based on NMF with orthogonal constraints, which had success in the classification of multi-omics data [8]. In the existing methods, multi-omics data are typically used for sample analysis, protein analysis or feature gene extraction analysis, and the co-expression network is rarely used for analysis and mining multi-omics information. The analysis of research based on networks built from multi-omics data may be more meaningful than networks built from a single type.

Based on the above problems, we discover that the series of NMF for integrative models have good performance for the reconstruction and mapping of multi-omics data. In view of previous work on NMF integrated models, we propose an integrated graph regularized non-negative matrix factorization model that is based on the perspective of gene network analysis and takes the spatial genometry of the genes into consideration. This model is similar to the above-mentioned integrated NMF models in form. However, the existing model is slow when dealing with large-scale data. In addition, current network models seldom consider the spatial geometric characteristics of the data, and most of them lack guidance for the construction of the gene co-expression network. We expand a Graph Regularized Non-negative Matrix Factorization (GNMF) [9] into an integrative model for multi-omics data network analysis, called iGMFNA. This model draws on the data reconstruction characteristics of NMF and the internal mapping advantages of graph regularization. We apply it to multi-omics data through matrix decomposition and iteration and provide a considerable foundation for the construction of gene expression networks. We confirm the effectiveness of the method by studying three types of multi-omics data for two cancers from TCGA: gene expression data (GE), copy number variation data (CNV), and methylation data (ME). The datasets from TCGA are high-dimensional with small samples, and we hope to identify or predict some cancer-related genes by establishing networks in large number genes. Firstly, we introduce iGMFNA to carry out data decomposition and fusion. Then, the gene co-expression networks are constructed by mapping variables matrix that corresponding three datasets. Finally, we can find some information related to cancers from the networks by text mining. In the aspect of network mining, to find more suspicious nodes, we combine the local and global properties of nodes of the networks.

The rest of this paper is organized as follows: Section II introduces related methods and Section III is the introduction for iGMFNA; the experiment results on two cancers data are given in Section IV; and Section V provides the conclusion of this paper.

II. RELATED WORKS

A. NMF

The original solution algorithm for NMF was proposed by Lee and Seung [10], [11]. NMF has extensive applications in image processing and bioinformatics. Given a non-negative data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, two non-negative matrices $\mathbf{W} \in \mathbb{R}^{m\times r}$ and $\mathbf{H} \in \mathbb{R}^{n\times r}(r << \min\{m, n\})$ can approximate it by minimizing the objective equation:

$$O_1 = \left\|\mathbf{X} - \mathbf{W}\mathbf{H}^{\mathbf{T}}\right\|^2, \tag{1}$$

where $\mathbf{X}$ is original cancer data matrix in our hypothetical experiment, m represents the number of genes (or samples) and n is the number of samples (or genes). r is a measure of dimensionality reduction. It can also be interpreted as the projection center of data $\mathbf{X}$ on $\mathbf{W}$ and $\mathbf{H}$. The NMF model can simply and effectively classify samples and identify differentially expressed genes that are widely accepted by researchers in bioinformatics.

B. GNMF

Compared with other improved NMF algorithms, NMF with manifold learning considers the internal spatial structure of the data and shows remarkable performance. The GNMF model is a classical improved NMF algorithm that incorporates manifold learning. The model minimizes the following equation:

$$O_2 = \left\|\mathbf{X} - \mathbf{W}\mathbf{H}^{\mathbf{T}}\right\|^2 + \lambda Tr\left(\mathbf{H}^{\mathbf{T}}\mathbf{L}\mathbf{H}\right), \tag{2}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{E}$ ($\mathbf{L} \in \mathbb{R}^{n\times n}$ or $\mathbb{R}^{m\times m}$, $\mathbf{D}_{ii} = \sum_j \mathbf{E}_{ij}$), $\mathbf{D}$ is diagonal matrix and the parameter $\lambda \geq 0$ controls the degree of smoothness of the equation. $\mathbf{L}$ is the Laplacian matrix, whose concept is based on spectral graph theory and manifold learning theory. Assume that the data vectors are distributed on a low-dimensional manifold embedded in a high-dimensional space: if two vectors x_i and x_j are close in the high-dimensional data space, then z_i and z_j are also close in the low-dimensional data space. Considering the neighbor geometry structure, we can easily find k nearest neighbors of every point x_i and assign the edges to them. Then, we can define the matrix of weight: $\mathbf{E}$. There are usually three ways to assign $\mathbf{E}$. The most commonly used and most effective way is 0-1 weight: $e_{ij} = 1$ if vertex i connects vertex j or $e_{ij} = 0$. We also redefine a new vector to indicate x_i on low-dimensional space, $\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2} \cdots\cdots, \mathbf{z_N}]^T$, inside $\mathbf{z_i} = [v_{i1}, v_{i2}, \cdots, v_{ik}]^T$, and European distance is defined as follows:

$$D\left(\mathbf{z_i}, \mathbf{z_j}\right) = \left\|\mathbf{z_i} - \mathbf{z_j}\right\|^2. \tag{3}$$

With the weight matrix, the initial equation can be written as:

$$\begin{aligned} R &= \frac{1}{2}\sum\nolimits_{i,j=1}^{N} \left\|\mathbf{z_i} - \mathbf{z_j}\right\|^2 \mathbf{E}_{ij} \\ &= \sum\nolimits_{i=1}^{N} \mathbf{z_i^T}\mathbf{z_i}\mathbf{D}_{ii} - \sum\nolimits_{i,j=1}^{N} \mathbf{z_i^T}\mathbf{z_j}\mathbf{E}_{ij} \\ &= Tr\left(\mathbf{H}^{\mathbf{T}}\mathbf{D}\mathbf{H}\right) - Tr\left(\mathbf{H}^{\mathbf{T}}\mathbf{E}\mathbf{H}\right) \\ &= Tr\left(\mathbf{H}^{\mathbf{T}}\mathbf{L}\mathbf{H}\right). \end{aligned} \tag{4}$$

This model considers the relationship between points, and it will make the data more modular.

III. METHOD

A. INTEGRATIVE FORMULA

The integrative model proposed in this paper is a data-driven model. It is designed for multi-omics data. Gene co-expression network analysis can increase the amount of prior knowledge available to the network if it can enhance intrinsic linkage of genes. Therefore, GNMF is introduced as the basis for the integrating model. The formula of the model can be written as follows:

$$O_3 = \sum_{I=1}^{d} \left\| \mathbf{X}_I - \mathbf{W}_I \mathbf{H}^{\mathbf{T}} \right\|^2 + \sum_{I=1}^{d} \lambda_I Tr(\mathbf{H}^{\mathbf{T}} \mathbf{L}_I \mathbf{H}) \quad s.t.\ \mathbf{W}_I \geq 0,\ \mathbf{H} \geq 0,\ I = 1, 2 \ldots, d, \tag{5}$$

where d is the number of datasets. One $\mathbf{X}$ corresponds to one $\mathbf{L}$, and the corresponding λ adjusts the smoothness of the matrix. The iterative formula of the corresponding variable is shown as follows:

$$(w_I)_{ij} \leftarrow (w_I)_{ij} \frac{(X_I H)_{ij}}{(W_I H^T H)_{ij}}, \quad h_{ij} \leftarrow h_{ij} \frac{\sum_{I=1}^{d} (X_I^T W_I + \lambda_I E_I H)_{ij}}{\sum_{I=1}^{d} (H_I W^T W + \lambda_I D_I H)_{ij}}. \tag{6}$$

$\mathbf{H}$ combines the information about different types of data in the iterative process, which is more conducive to excavating information and analyzing data theoretically. For the data matrix $\mathbf{X}$, the decomposition can get the matrix $\mathbf{H}$ related to the sample. If it is transposed, it will decompose to obtain the matrix $\mathbf{H}$ related to the gene. In the parameter test, we would conduct the sample clustering to obtain sample information, and when constructing the network, we want get the information about genes to transpose $\mathbf{X}$. In the experiments, we can obtain a sample-related fusion matrix or a gene-related fusion matrix $\mathbf{H}$ by transposing the data $\mathbf{X}$, and then carrying out the network analysis.

B. NETWORK CONSTRUCTION AND MINING

We use the Pearson correlation coefficient (PCC) to measure the relationships of nodes in networks to obtain the adjacency matrix, and then sort absolute values of the PCC matrix and perform polynomial curve fitting. Finally, we choose the first inflection point as the filtering threshold of the matrix to obtain final networks.

Node mining is a key step in network mining. We want to find genes that may affect the entire network. When these genes are abnormally expressed, they are highly related to cancer. Therefore, we must consider some specific attributes of each node. A node has many features and it is difficult to determine which one is most prominent. Starting from the local and global characteristics of the nodes, we hope to better evaluate the criticality of a node in the network by combining some pivotal features. After evaluating many measures, the three most commonly applied are degree, betweenness and closeness. To improve the discovery of abnormal nodes (indicating genes that are associated with pathogenesis), we define a multi-measure Score for every node x:

$$Score(x) = D \times B / C, \tag{7}$$

where D is degree of x, B is betweenness of x and C is closeness of x. Degree represents a local property of the node in the network, and the betweenness and closeness reflect global properties of the node in the network, so this metric combines both global and local features of a node. By examining the Score for every gene in the network, we can discriminate which genes are most suspicious. The analysis of iGMFNA is briefly shown in Figure 1. There are two major steps with in iGMFNA: the first step is the process of integrative GNMF and the second step is network construction based on the fusion matrix and network mining.

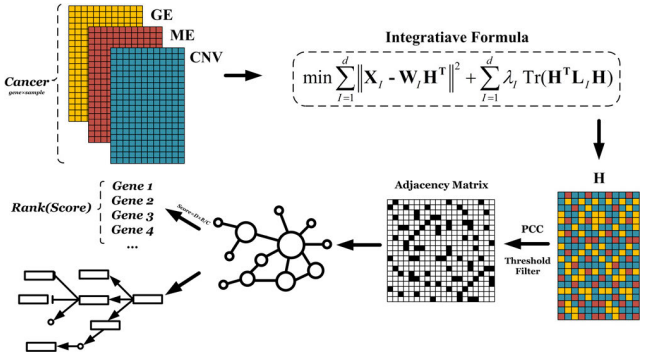


FIGURE 1. This is the schematic diagram for network analysis of iGMFNA. There are 3 types of datasets for cancer in this paper. The fusion matrix can be calculated by integrative formula and build the networks. Through the combination of different properties of nodes, some suspicious genes and pathways can be detected.

IV. RESULTS

A. DATASETS

There are two datasets of cancers used in this study. They are all multi-omics data: Cholangiocarcinoma (CHOL) and Pancreatic adenocarcinoma (PAAD), obtained from TCGA. Both cancer datasets contain three types of data: GE, ME, and CNV. The integrated model in our experiments is a three-dimensional integrated model ($\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$). All three types of data were collected from the same samples and the genes have been aligned experimentally. The status of the samples covers two types: tumor or normal. CHOL contains 19876 genes in 45 samples (36 positive samples). PAAD contains 19877 genes in 180 samples (176 positive samples). Some datasets, such as CNV and ME, measured less expression values of genes than GE. Therefore, for different types of the same cancer, we retain the common genes of every datasets to keep the scale of the input matrices consistent. The datasets are summarized as listed in Table 1. The ME data contains negative values, so we normalized these data to maintain suitable input for the model.

B. MODELS COMPARISON

For the three-dimensional integrative model, the parameters we need to control are the dimension reduction parameter

TABLE 1. Summary of dataset.

Cancer	Data Types	Dimensions (gene × sample)	Sample status
CHOL	GE ME CNV	19876×45	Tumor or Normal
PAAD	GE ME CNV	19877×180	

r and the joining manifold learning parameter λ. In particular, λ corresponds to three datasets with three variables $(\lambda_1, \lambda_2, \lambda_3)$. To obtain better parameters and ensure the reliability of the decomposed matrix, we use the sample-related fusion matrix $\mathbf{H}$ that is ready for network construction to test the effect of iGMFNA model. We determine λ and r by comparing the clustering effect on the sample-related fusion matrix $\mathbf{H}$. The clustering effects are assessed by contrasting clustered labels with true label. The accuracy (ACC) [12] as evaluation function:

$$ACC = \frac{\sum_{i=1}^{n} \delta\left(s_i, \text{map}\left(l_i\right)\right)}{n}, \tag{8}$$

where n is the total number of samples, $\delta(x, y)$ is a delta function: if $x = y$, $\delta(x, y) = 1$; otherwise, $\delta(x, y) = 0$. $map\left(l_i\right)$ is mapping function that maps each cluster label l_i to original label s_i. Surprisingly, the performances in the experiments are more stable and effective on all datasets when λ_I is 0.01 (this is the experimental experience value of this paper, not unique to other datasets). Compared with parameter λ, the dimension reduction parameter r has a greater impact. Theoretically, the larger r can obtain the better recovery of data. However, the principle of NMF is to restore the data best by setting the smaller r ($r << \{m, n\}$) to achieve the purpose of dimension reduction. We want to get the smaller r when the clustering effect is guaranteed, and conduct a number of experiments on every model to select the better parameter r, as shown in the Figure 2. As can be seen from Figure 2, as r increases, the clustering accuracies of the four models increase overall, especially iGMFNA, the improvement is relatively stable. The CHOL in Figure 2(a) is more separable and the overall effects of models are considerable. When $r = 5$, the performances of four models are acceptable, and it is also a stable rising point; the performances of models on PAAD in Figure 2(b) are more unstable, when $r = 40$, it is the point where the four models have the largest increase.

To prove the validity of our method, we compare several similar types of integrative models. The jNMF model integrates the basic NMF into an integrated model; iONMF and iNMF are models that are based on orthogonal constraints and sparse norm constraints, as listed in Table 2.

The results in Table 2 are the average clustering accuracy (ACC_Mean) of 50 runs of each model. The first value is mean value after 50 times, and the second value is the standard deviation (the squared value of variance). The setting of the dimension reduction parameter r is unified across all

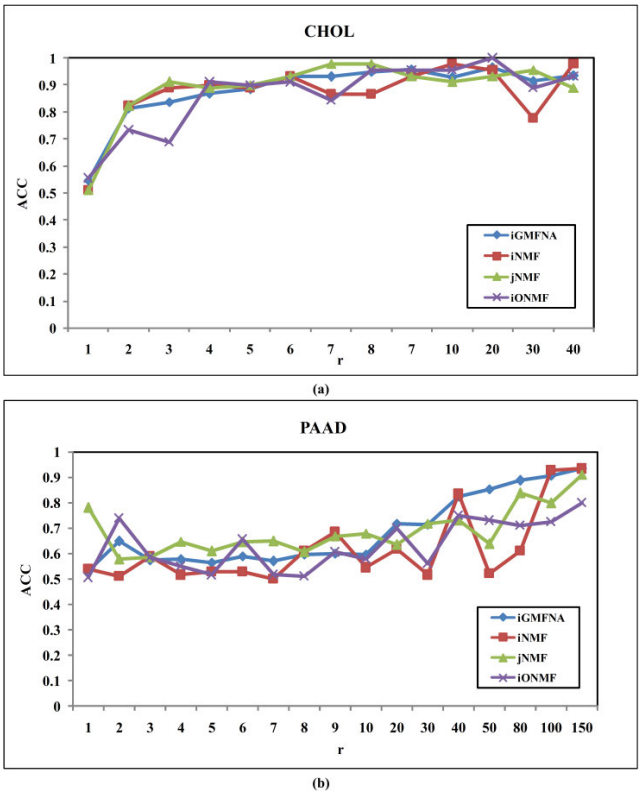


FIGURE 2. Selection of parameter *r*. For NMF, we chose a smaller *r* in the better case of data recovery. When *r* = 5 and 40, the effect of each model is good.

TABLE 2. Comparison of clustering accuracy for integrative models.

Model	ACC_Mean (50 times)	
	CHOL (r=5)	PAAD (r=40)
iGMFNA	**0.95±0.0905**	**0.87±0.1520**
iNMF	0.91±0.0501	0.86±0.1595
iONMF	0.83±0.0690	0.75±0.1800
jNMF	0.92±0.1194	0.86±0.1561

models for each dataset. As observed in Table 2, iGMFNA is better able to reconstruct the CHOL dataset compared to other integrative models. Since the CHOL dataset is more separable, all models have good results. With the PAAD dataset, the difference in the effect of these models is not particularly obvious. Further comparisons can better understand the effects of these models.

Using the PAAD dataset as an example (the convergence on CHOL is same as PAAD), the convergence of the four models is presented in Figure 3. The error value [6] is defined as follows:

$$Error = \sum_{I=1}^{d} \frac{mean(mean(\left|\mathbf{X}_I - \mathbf{W}_I\mathbf{H}^T\right|))}{mean(mean(\mathbf{X}_I))}, \tag{9}$$

where mean$(\cdot)$ is the mean function (the average of all the numbers in a matrix). The convergence of iGMFNA is relatively stable. Although the convergence of iONMF is

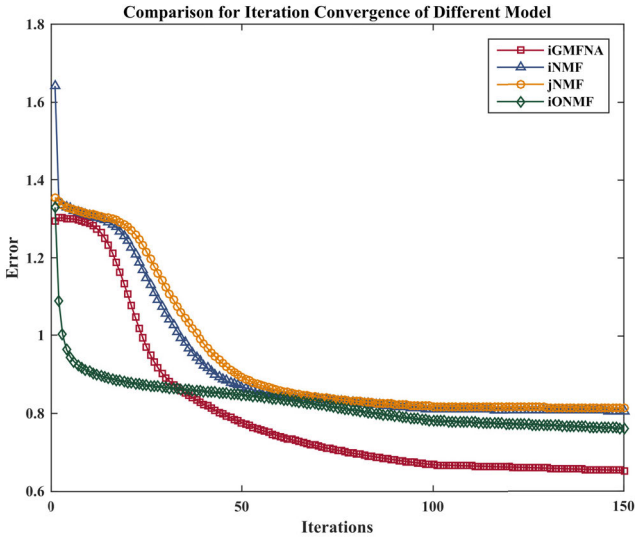


FIGURE 3. Comparison of convergence about four models. The convergence effect of every model is considerable; the convergence speed is still relatively fast. Among them, iONMF is the fastest, but overall, iGMFNA has the best convergence effect. And the convergence effect is still partially different.

TABLE 3. Run time comparison.

Models	iGMFNA	iNMF	iONMF	jNMF
Time(Second)	4619.3649	10406	81614	10317

fast, the run time of iONMF is spent too much, and the later convergence effect is inferior to iGMFNA. We also compare the runtime of these models, as listed in Table 3 (on 64-bit win 7 operating system with 64GB RAM, i7-6700 CPU, 3.40G HZ); the values of table 3 are the runtime comparison of the above integrative models with the same number of iterations. The smaller the runtime is, the faster the model runs. iGMFNA is superior in run time. In fact, this process is also related to the solution of the single model. The solution process of the integrated model relies mainly on the solution of the single model, so the choice of single model becomes important in the calculation.

Based on these experimental effects, we can consider that the fusion matrix **H** related with genes based on this model is reliable for network construction.

C. NETWORK ANALYSIS

For the two cancer datasets, we set 20 nodes as the baseline to reserve 8 modules for every cancer (CHOL: 698 nodes, 1506 edges; PAAD: 542 nodes, 3211 edges). Modules are visualized by Cytoscape [13] in Figure 4 and Figure 5. Larger node size in the figure is related to a higher degree of connectivity and darker color corresponds to larger betweenness. Every node in the module has its own distinctive Score. The node with the greatest Score is the object we will excavate. In Figure 4, we display a major module based on CHOL.

TABLE 4. Genes with higther score on CHOL networks.

Gene	Comments
SMEK3P	Protein Coding gene
OR8H1	Among its related pathways are Signaling by GPCR and Olfactory Signaling Pathway.
IL31	IL31 may be involved in the promotion of allergic skin disorders and in regulating other allergic diseases, such as asthma.
DDX53	Diseases associated with DDX53 include Non-Secretory Myeloma and Gastrointestinal System Cancer.
CXCL13	Diseases associated with CXCL13 include Angioimmunoblastic T-Cell Lymphoma and Burkitt Lymphoma.
PAX4	Diseases associated with PAX4 include Maturity-Onset Diabetes Of The Young, Type 9 and Diabetes Mellitus, Ketosis-Prone.
PRR35	Protein Coding gene
USP17L9P	Among its related pathways are Ubiquitin-Proteasome Dependent Proteolysis.
GAGE12D	Protein Coding gene
TNNT3	Diseases associated with TNNT3 include Arthrogryposis, Distal, Type 5 and Digitotalar Dysmorphism.

TABLE 5. Genes with higther score on PAAD networks.

Gene	Comments
OR5W2	Among its related pathways are Signaling by GPCR and Olfactory Signaling Pathway.
POPDC3	This gene is expressed in cardiac and skeletal muscle and may play an important role in these tissues during development.
SLC25A14	Among its related pathways are Metabolism and Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
ARMCX3	Aliases for Alex3, this gene encode a member of the ALEX family of proteins which may play a role in tumor suppression.
PCDHGA3	This gene is a member of the protocadherin gamma gene cluster, one of three related clusters tandemly linked on chromosome five. These gene clusters have an immunoglobulin-like organization, suggesting that a novel mechanism may be involved in their regulation and expression.
OR9G4	Among its related pathways are Signaling by GPCR and Olfactory Signaling Pathway.
PDK3	Diseases associated with PDK3 include Charcot-Marie-Tooth Disease, X-Linked Dominant, 6 and X-Linked Charcot-Marie-Tooth Disease.
HTATSF1	Protein Coding gene.
IVL	Diseases associated with IVL include Porokeratosis and Skin Disease.
CETN2	Diseases associated with CETN2 include Xeroderma Pigmentosum, Complementation Group C.

This module contains the most nodes and can be considered as the most productive module. The genes (nodes) extracted from the modules according to their Scores can be as suspicious genes. We refer to the annotations on GeneCards (https://www.genecards.org/) for the top 10 genes, listed in Table 4.

Many cancers-related genes have been discovered in the top 10 nodes. For example, research has indicated that polymorphisms of IL31 are linked with bladder cancer [14]; and

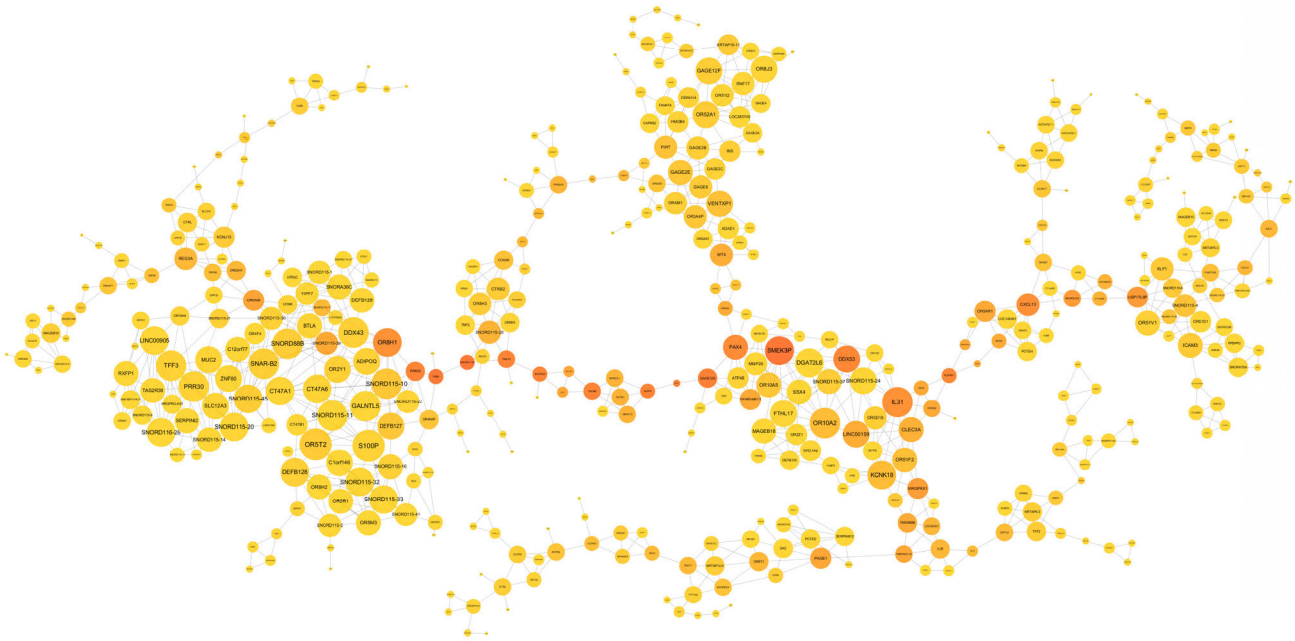


FIGURE 4. Network modules construction based on CHOL. There is a major module. According to the definition of node evaluation, we find top 10 abnormal genes such as SMEK3P, OR8H1, IL31, DDX53, CXCL13, PAX4 and so on. Among them, IL31, DDX53, CXCL13 and PAX4 are associated with cancers. Some genes have significant values on single measure, however, no cancerous records were found.

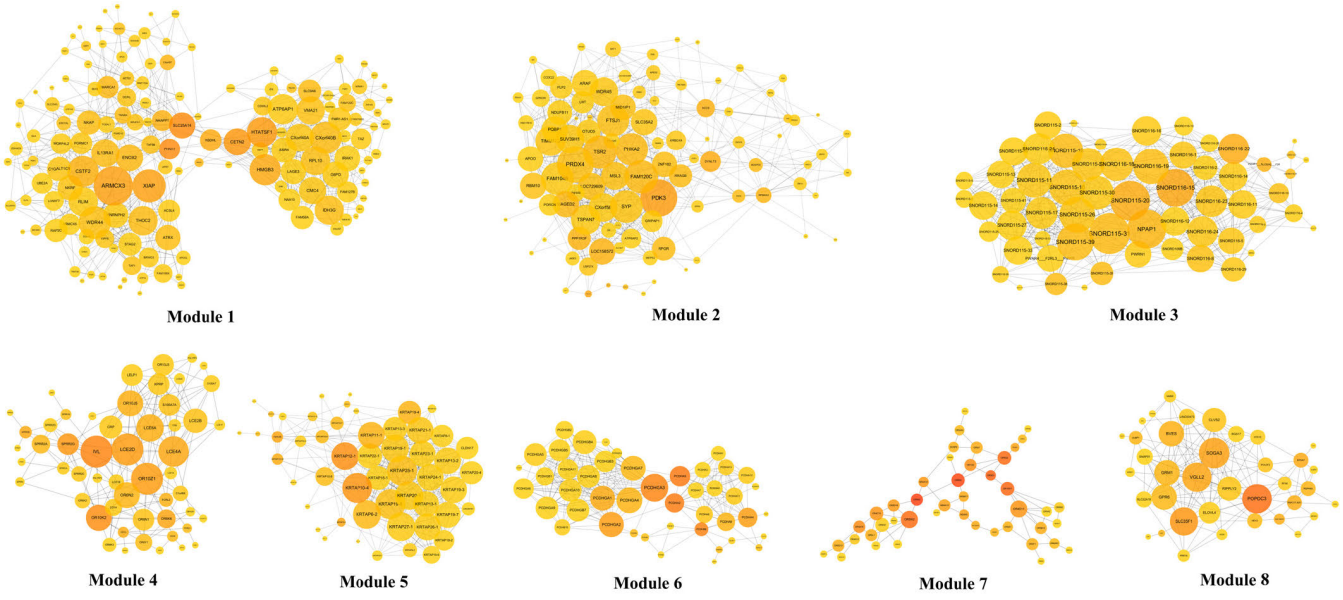


FIGURE 5. Network modules construction based on PAAD. We also keep 8 major modules. There are 5 genes from top 10 genes that are verified with cancers (POPDC3, ARMCX3, PCDHGA3, PDK3, CETN2). Among them, CETN2 was validated in PAAD. Some genes with higher Scores have not been clinically confirmed, such as OR5W2 and SLC25A14; they could be potential research targets.

IL31 has a potential role in the prognosis of endometrial cancer patients [15]. DDX53 confers resistance to anti-cancer drugs in breast cancer cells [16]. CXCL13 was shown to be over-expressed in breast cancer tissues [17]; concentration of CXCL13 in cerebrospinal fluid has also been correlated with the degree of blood-brain barrier disruption in lymphoma cases where malignant lymphocytes that have targeted the central nervous system (CNS) [18]. Zhang et al. found that the expression of PAX4 was dysregulated in a variety of human cancers and considered that it may be important in multiple tumors as a driver gene [19].

The modules reserved for PAAD are smaller in size, as shown in Figure 5. We extracted some of the higher-Score genes from these modules, as listed in Table 5.

TABLE 6. Pathways detected by networks of different cancers.

CHOL			PAAD		
GeneSet	P-value	FDR	GeneSet	P-value	FDR
GPCR downstream signaling(R)	1.11E-16	2.22E-14	Cadherin signaling pathway(P)	1.11E-16	2.56E-14
Olfactory transduction(K)	1.11E-16	2.22E-14	Olfactory transduction(K)	1.11E-16	2.56E-14
Keratinization(R)	1.40E-09	1.86E-07	**Wnt signaling pathway(P)**	1.82E-10	2.80E-08
GPCR ligand binding(R)	1.25E-07	1.25E-05	GPCR downstream signaling(R)	2.36E-09	2.72E-07

The literature has reported that POPDC3 promoter regions were hypermethylated in the gastric cancer cell lines where they were silenced [20]. Du et al. discovered that reduction of ARMCX3 was correlated with the development of non-small cell lung cancer [21]. In Wilms' tumor, hypermethylation of PCDHs including PCDHGA3 led to gene silencing, and β-catenin protein was elevated, promoting the activity of β-catenin/T-cell factor (TCF) reporter activity and the Wnt signaling pathway [22]. PDK3 is a potential target for mitochondrial modulation in colorectal cancer [23]. CETN2 and seven other genes were validated in PAAD in [24]. Some genes with higher Scores have not been clinically confirmed, including SMEK3P, OR8H, OR5W2 and SLC25A14; their potential significance values need further clinical study.

In the networks, we also match some pathways (by KEGG and Reactome), which with the smaller p-values, as listed in Table 6. The p-value and FDR are given by Reactome FI in Cytoscape plugin when carrying out pathway enrichment analysis, based on the fisher's exact test. When the p-value is larger, the evidence against the null hypothesis is weaker. Conversely, the smaller the p-value is, the stronger the evidence against the null hypothesis. Related studies have shown that Wnt signaling pathway is closely related to PAAD [25], [26]; in fact, it is a pathway closely associated with many other cancers such as breast, lung, ovarian, and colorectal.

D. DISCUSSION

The generation of multiple types of data makes the use of an integrated model (sometimes also called a multi-view model) widely proposed in recent years. Among many machine learning methods, a series of models based on NMF have always performed well in data recovery and reconstruction [27]. For the analysis and construction of the integration model, we have achieved some improved results in the fusion matrix obtained by iGMFNA. The integrated model-iGMFNA, which incorporates manifold learning concepts, takes the internal geometry of every type of data into account, allowing the heterogeneity of the data to be captured, so that the decomposed matrix maintains the internal relationships of multi-omics data. Building networks on a fusion matrix with multi-source information is also conducive to detecting more information about cancer.

Considering the node mining of the networks, there are many node properties and each property has a reasonable interpretability. Within a biological network, it is difficult to estimate which property can distinguish abnormally expressed genes well, and it is difficult to have a unified measure for mining nodes in the network. Since some abnormal nodes may not be prominent when using a single feature. In this case, it is necessary to combine features in a rational way to achieve better outcomes, so we define a new scoring measure by combining several important node features. The measurement of node mining (Score) is also obtained by a lot of experimental tests and comparisons. When mining suspicious nodes in the network, we consider the local and global characteristics of each gene so that every gene is subject to a relatively comprehensive assessment, leading to the discovery of more valuable nodes. For example, in Figure 4, IL31, whose abnormal expression was confirmed to be associated with multi-cancers, was not the most prominent on the three scales individually (degree, betweenness and closeness), but it could be detected under the multi-measure Score.

Through the iGMFNA model, we found many genes that are associated with many cancers. We speculate that the pathogenesis of the two cancers in this study is likely to be related to other cancers, and the development of PAAD and CHOL can be accompanied by the development of other cancers. Of course, iGMFNA can also be applied for the mining and analysis of other cancer or disease data, and the model can be continuously expanded with the increase of available multi-omics data.

V. CONCLUSION

This paper proposes a network analysis method based on the integrated model, iGMFNA. By introducing manifold learning into the integrative model, spatial geometry of different types of data is detected and the construction of networks for node relationships is improved. When compared with similar types of integrated NMF models, iGMFNA is still superior in data fusion and reconstruction. In terms of node mining, we comprehensively analyze and combine several different important properties of nodes in the network to define a Score with local and global characteristics for each node for identification of suspicious genes. Based on the networks constructed from the fusion matrix, we have identified some higher-Score genes and pathways that are closely related to the prognosis and pathogenesis of multi-cancers, and it can be inferred that these genes may have similar effects in the two cancers we analyzed. Other genes detected in this paper are also worthy of further clinical validation and analysis.

ACKNOWLEDGMENT

The authors would like to thank for the reviewers' comments concerning our before manuscript, those comments are all

valuable and very helpful for revising and improving our article.

REFERENCES

[1] S.-G. Ge, J. Xia, W. Sha, and C.-H. Zheng, "Cancer subtype discovery based on integrative model of multigenomic data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 5, pp. 1115–1121, Sep./Oct. 2017.

[2] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1, pp. A68–A77, Jan. 2015.

[3] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, Jan. 2014.

[4] S. Patkar, A. Magen, R. Sharan, and S. Hannenhalli, "A network diffusion approach to inferring sample-specific function reveals functional changes associated with breast cancer," *PLoS Comput. Biol.*, vol. 13, no. 11, Nov. 2017, Art. no. e1005793.

[5] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, Nov. 2009.

[6] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Res.*, vol. 40, no. 19, pp. 9379–9391, Oct. 2012.

[7] Z. Yang and G. Michailidis, "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data," *Bioinformatics*, vol. 32, no. 1, pp. 325–342, Jan. 2015.

[8] M. Stražar, M. Žitnik, B. Zupan, J. Ule, and T. Curk, "Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins," *Bioinformatics*, vol. 32, no. 10, pp. 1527–1535, May 2016.

[9] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001, pp. 556–562.

[12] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2003, pp. 267–273.

[13] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, 2003.

[14] Q. Li, T. Tang, P. Zhang, C. Liu, Y. Pu, Y. Zhang, H. Song, Y. Wang, Y. Song, M. Su, and B. Zhou, "Correlation of IL-31 gene polymorphisms with susceptibility and clinical recurrence of bladder cancer," *Familial Cancer*, vol. 17, no. 4, pp. 577–585, Oct. 2018.

[15] X. Zeng, Z. Zhang, Q.-Q. Gao, Y.-Y. Wang, X.-Z. Yu, B. Zhou, and M.-R. Xi, "Clinical significance of serum interleukin-31 and interleukin-33 levels in patients of endometrial cancer: A case control study," *Disease Markers*, vol. 2016, pp. 1–7, May 2016.

[16] H. Kim, Y. Kim, and D. Jeoung, "DDX53 promotes cancer stem cell-like properties and autophagy," *Molecules Cells*, vol. 40, no. 1, pp. 54–65, Jan. 2017.

[17] J. Panse, K. Friedrichs, A. Marx, Y. Hildebrandt, T. Luetkens, K. Bartels, C. Horn, T. Stahl, Y. Cao, K. Milde-Langosch, A. Niendorf, N. Kröger, S. Wenzel, R. Leuwer, C. Bokemeyer, S. Hegewisch-Becker, and D. Atanackovic, "Chemokine CXCL13 is overexpressed in the Tumour tissue and in the peripheral blood of breast cancer patients," *Brit. J. Cancer*, vol. 99, no. 6, pp. 930–938, Sep. 2008.

[18] L. Fischer, A. Korfel, S. Pfeiffer, P. Kiewe, H.-D. Volk, H. Cakiroglu, T. Widmann, and E. Thiel, "CXCL13 and CXCL12 in central nervous system lymphoma patients," *Clin. Cancer Res.*, vol. 15, no. 19, pp. 5968–5973, Oct. 2009.

[19] J. Zhang, X. Qin, Q. Sun, H. Guo, X. Wu, F. Xie, Q. Xu, M. Yan, J. Liu, Z. Han, and W. Chen, "Transcriptional control of PAX4-regulated miR-144/451 modulates metastasis by suppressing ADAMs expression," *Oncogene*, vol. 34, no. 25, pp. 3283–3295, Aug. 2014.

[20] M. Kim, H.-R. Jang, K. Haam, T.-W. Kang, J.-H. Kim, S.-Y. Kim, S.-M. Noh, K.-S. Song, J.-S. Cho, H.-Y. Jeong, J. C. Kim, H.-S. Yoo, and Y. S. Kim, "Frequent silencing of popeye domain-containing genes, BVES and POPDC3, is associated with promoter hypermethylation in gastric cancer," *Carcinogenesis*, vol. 31, no. 9, pp. 1685–1693, Sep. 2010.

[21] J. Du, X. Zhang, H. Zhou, Y. Miao, Y. Han, Q. Han, and E. Wang, "Alex3 suppresses non–small cell lung cancer invasion via AKT/Slug/E-cadherin pathway," *Tumour Biol.*, vol. 39, no. 7, Jul. 2017, Art. no. 1010428317701441.

[22] A. R. Dallosso, A. L. Hancock, M. Szemes, K. Moorwood, L. Chilukamarri, H.-H. Tsai, A. Sarkar, J. Barasch, R. Vuononvirta, C. Jones, K. Pritchard-Jones, B. Royer-Pokora, S. B. Lee, C. Owen, S. Malik, Y. Feng, M. Frank, A. Ward, K. W. Brown, and K. Malik, "Frequent long-range epigenetic silencing of protocadherin gene clusters on chromosome 5q31 in wilms' tumor," *PLoS Genet.*, vol. 5, no. 11, Nov. 2009, Art. no. e1000745.

[23] S. Yeluri, B. Madhok, P. Prasad, M. J. Alemkunnapuzzha, H. L. Thorpe, S. L. Perry, T. A. Hughes, K. R. Prasad, P. Quirke, and D. G. Jayn, "Exploiting warburg's effect for clinical gain: Pdk-3 is a potential target for mitochondrial modulation in colorectal cancer (CRC)," *Int. J. Surgery*, vol. 8, no. 7, p. 514, Jan. 2010.

[24] H. He, Y. Di, M. Liang, F. Yang, L. Yao, S. Hao, J. Li, Y. Jiang, C. Jin, and D. Fu, "The micro RNA-218 and ROBO-1 signaling axis correlates with the lymphatic metastasis of pancreatic cancer," *Oncol. Rep.*, vol. 30, no. 2, pp. 651–658, Jun. 2013.

[25] B. Garg, B. Giri, K. Majumder, V. Dudeja, S. Banerjee, and A. Saluja, "Modulation of post-translational modifications in β-catenin and LRP6 inhibits Wnt signaling pathway in pancreatic cancer," *Cancer Lett.*, vol. 388, pp. 64–72, Mar. 2017.

[26] Z. Wang, Q. Liu, H. Yu, L. Zhao, S. Shen, and J. Yao, "Blockade of SDF-1/CXCR4 signalling inhibits pancreatic cancer progression *in vitro* via inactivation of canonical Wnt pathway," *Brit. J. Cancer*, vol. 99, no. 10, pp. 1695–1703, Oct. 2008.

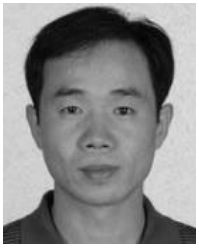
[27] J.-X. Liu, D. Wang, Y.-L. Gao, C.-H. Zheng, Y. Xu, and J. Yu, "Regularized non-negative matrix factorization for identifying differentially expressed genes and clustering samples: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 3, pp. 974–987, May/Jun. 2018.

YING-LIAN GAO received the B.S. and M.S. degrees from Qufu Normal University, China, in 1997 and 2000, respectively.

She is currently a Lecturer with the Qufu Normal University Library. Her current research interests include data mining and pattern recognition.

MI-XIAO HOU received the B.S. and M.S. degrees in computer science and technology from QuFu Normal University, China, in 2016 and 2019, respectively.

She is currently pursuing the Ph.D. degree in computer science and technology with the Harbin Institute of Technology, Shenzhen, China. Her research interests include pattern recognition and bioinformatics.



JIN-XING LIU (M'12) received the B.S. degree in electronic information and electrical engineering from Shandong University, China, in 1993, the M.S. degree in control theory and control engineering from Qufu Normal University, China, in 2003, and the Ph.D. degree in computer simulation and control from the South China University of Technology, China, in 2008.

From June 2011 to December 2015, he was with Shenzhen Graduate School, Harbin Institute of Technology, as a Postdoctoral Research Fellow. He is currently a Professor with the School of Information Science and Engineering, Qufu Normal University. His research interests include pattern recognition, machine learning, and bioinformatics.

XIANG-ZHEN KONG received the B.S. degree in computer science and technology and the M.S. degree in control theory and control engineering from Qufu Normal University, China, in 2002 2008, respectively, where she is currently an Associate Professor with the School of Information Science and Engineering. Her research interests include pattern recognition and bioinformatics.