NIPMI: A Network Method Based on Interaction Part Mutual Information to Detect Characteristic Genes from Integrated Data on Multi-Cancers

QIAN DING1, JUNLIANG SHANG1,2, Member, IEEE, YAN SUN1, GUANGSHUAI LIU3, FENG LI1, XIGUO YUAN4, JIN-XING LIU1, Member, IEEE
1 School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China
2 School of Statistics, Qufu Normal University, Qufu 273165, China
3 College of Life Science, Qufu Normal University, Qufu 273165, China
4 School of Computer Science and Technology, Xidian University, Xi’an 710071, China

Corresponding author: Junliang Shang (e-mail: shangjunliang110@163.com).

This work was supported by the National Science Foundation of China (61972226, 31900311, 61902216, 61571341, 61572284) and by the China Postdoctoral Science Foundation (2018M642635).

ABSTRACT Comprehensive analysis of integrated data on multi-cancers is important for understanding the biological mechanism of these cancers at the system level. Network methods give us new insight into simultaneously identifying the characteristic genes and pathways from multi-cancers. Nevertheless, when measuring the similarity quantification of genes, it is a challenge to choose suitable methods for network construction and analysis. Herein, the NIPMI method, based on Interaction Part Mutual Information (IPMI) measure for detecting characteristic genes from multi-cancers data, is proposed. First, Robust PCA was applied to select genes for network construction. Then, a network construction measure, IPMI, was proposed to effectively quantify the similarity between genes in the network, which is the highlight of NIPMI. Furthermore, we introduced a novel topological property, Topological Score, that combined the local and global properties of each node to find more candidate nodes in the network. Finally, pathway enrichment analysis was performed to validate the biological functions of multi-cancers. The experimental results demonstrated that NIPMI facilitates the identification of characteristic genes in a multicancer network; thus, it may serve as a valuable tool for detecting characteristic genes and significantly enriched pathway terms.

INDEX TERMS Multi-cancers, Interaction part mutual information (IPMI); Gene interaction network, Characteristic genes.

I. INTRODUCTION

The study of cancer diagnostics and therapeutics has traditionally been a hot topic [1]. With the rapid development of DNA sequencing technology and the completion of the human genome project, a large amount of data on multi-cancers has been generated. The Cancer Genome Atlas (TCGA) project systematically collected a large amount of such data [2]. According to recent studies, disparate cancers have many shared oncogenic mutations, and some characteristic genes that are not highly frequently mutated in specific cancers may show considerable mutation frequencies in patients with multi-cancers. Consequently, identifying characteristic genes that are relevant to cancer progression from multi-cancers data is an important task in understanding tumorigenesis [3].

Various methods for detecting the characteristic genes associated with multi-cancers have been proposed. Kaczkowski et al. [4] used a genome-wide expression profiling approach to identify a comprehensive set of candidate biomarkers with multicancer potential. Huang et al. [5] explored the biomarkers of multi-cancers using a feature selection method based on gene essentiality. In addition, because research at the level of the function of a single gene has limited the scope and process of exploring the biological functions of cancer cells, detection of characteristic genes has gradually shifted from the individual gene level to the network level [6, 7]. Exploring characteristic genes and pathogenic pathways based on gene networks constructed using multi-cancers data will provide an unprecedented opportunity to discover the molecular mechanisms of cancer [8-10]. Recently, Hou et al. [7] combined the network and low-rank methods and used the combined method to mine information from integrated data on multi-cancers after noise reduction. Park...
et al. [11] proposed an NTri-Path method to find cancer-type-specific pathways using mutation data from multiple types of cancers based on network analysis. Sun et al. [12] developed the CoFu method to identify the commonalities and differences in multi-cancers omics markers using a novel penalization technique.

In a gene network, the presence of edges between two nodes (genes) indicates that the two nodes are associated with each other; such edges are usually identified by calculating the similarity between genes. Many measures for quantifying similarities between genes are currently available. Among them, the Pearson correlation coefficient (PCC) is the method most commonly used to quantify the linear correlation between two genes [13, 14]. However, distinguishing indirect and direct associations between genes using PCC is difficult; furthermore, the method also has an overestimation problem [15]. The Spearman correlation coefficient (SCC) also quantifies the linear correlation in the gene interaction network [16], but its robustness is worse than that of PCC. To make the distribution of the PCC conform to the non-scale standard, the addition of an index based on PCC is proposed when a WGCNA package is used to construct a network, but the index is not unique [17]. Mutual information (MI), which is used to quantify the nonlinear correlation between two genes, cannot detect direct dependencies, and it also has an overestimation problem [18]. The maximum information coefficient (MIC) is used to detect linear and nonlinear direct correlations between genes. However, it is still controversial whether the MIC or the MI is more suitable for measuring the interactions between genes [19, 20]. The co-information (CI) measure considers that CI and MI values for a single gene and for a class of genes are both positive [21] and that they are equivalent; the values of CI and MI between two or more genes and classes are not equivalent, while the values of CI may be positive or negative [22]. Conditional mutual information (CMI), which is superior to linear measures, is used to quantify the nonlinear direct correlation between genes. However, for genes that are tightly connected in the network, CMI is subject to false negative or underestimation problems [23].

Part mutual information (PMI) [24] is used to quantify the nonlinear direct correlation between variables. PMI defines partial independence, and its value is always greater than or equal to the value of CMI. However, PMI is not proportional to the direct association, although it may overcome the underestimation of CMI. All in all, due to the limited suitability of each of these measures for effectively quantifying the similarity between nodes in a network, selection of a suitable network construction measure is still a challenge.

This paper proposes a novel network construction and analysis method, NIPMI, based on the Interaction Part Mutual Information (IPMI) measure, for the detection of characteristic genes. In this method, the Robust PCA feature selection method is first used to screen network construction genes from integrated multi-cancers data sets.

Then, to overcome the disadvantages of the current network construction measures, we propose the use of a new IPMI measure to construct a gene interaction network. In addition to quantifying nonlinear direct dependencies between genes, IPMI can also adjust for the interactions between different genes according to information on whether a given pair of genes is redundant or interactive. Third, we used a new Topological Score (TS) that combines the local and global properties of nodes to detect the characteristic genes in the networks. Lastly, biological analyses of the network construction genes were performed to identify enriched pathways. Based on the experimental results presented in this paper, we speculate that these characteristic genes and enriched pathways are likely to be potential factors in the development and progression of these cancers.

II. METHOD

A. SELECTION OF GENES FOR NETWORK CONSTRUCTION USING ROBUST PCA

Consider a matrix $M$ of integrated gene expression data with size $m \times n$. The matrix is given by $M = S + A$; $m$ represents the number of genes, and $n$ is the number of samples in matrix $M$. Since Robust PCA (RPCA) pays particular attention to the underlying class structure, it is always an optimal dimensionality reduction procedure for feature selection purposes. To reduce the difficulty of gene interaction network construction, we used RPCA to select genes for network construction. RPCA solves the following optimization problem:

$$
\text{min} \| A \|_* + \lambda \| S \|_* \text{, subject to } M = A + S
$$

where $\| A \|_* = \sum \sigma_i(A)$ represents the nuclear norm of the matrix $A$, $\| S \|_* = \sum \| S_{ij} \|$ represents the $L_1$-norm of $S$, $\lambda$ is a positive regulation parameter $\lambda = (c * \max(m, n))^{1/2}$, and $c$ is set to 0.3 according to [25]. To solve the RPCA problem, we introduced the inexact Augmented Lagrange Multipliers ALM (IALM) method [26]; equality constraints were eliminated by introducing a Lagrange multiplier. We can then obtain the sparse perturbation matrix $S$. Therefore, the differentially expressed genes can be determined by analyzing the sparse matrix $S$:

$$
S = \begin{pmatrix}
S_{11} & \cdots & S_{1n} \\
\vdots & \ddots & \vdots \\
S_{m1} & \cdots & S_{mn}
\end{pmatrix}
$$

We obtain $S_0$ by summing the absolute values of the factors in $S$ by line:

$$
S_0 = \left[ \sum_{i=1}^n |S_{i1}| \ldots \sum_{i=1}^n |S_{in}| \right]^T
$$

Based on the results of sorting $S_0$, we select the value corresponding to the point at which the difference in slope is the smallest as the number of differentially expressed genes.
B. CONSTRUCTION OF A GENE INTERACTION NETWORK

1) INFORMATION THEORETIC NOTIONS

Since information entropy provides a way to quantify the information encoded in discrete variables, we first approximate the continuous distribution of gene expression data using a finite discrete distribution. Given a gene $X$, $H(X)$ is the measure of the average uncertainty of gene $X$:

$$ H(X) = - \sum_{x \in X} p(x) \log p(x) $$

(4)

where $p(x) = \Pr(X = x)$ is the probability of each gene $X$. The joint entropy $H(X, Y)$ of two genes $X$ and $Y$ is defined as

$$ H(X, Y) = - \sum_{x \in X, y \in Y} p(x, y) \log p(x, y) $$

(5)

where $p(x, y)$ is the joint probability of $X$ and $Y$: $p(x, y) = \Pr(X = x, Y = y)$, $x \in X$ and $y \in Y$.

The mutual dependence commonly found between two genes $X$ and $Y$ is of significance, and this is quantified by the mutual information (MI). Based on Eq (4) and Eq (5), the definition of MI is as follows:

$$ MI(X, Y) = H(X) + H(Y) - H(X, Y) = H(X, Y) - H(X | Y) $$

(6)

Based on the notion of mutual information, Jakulin et al. proposed interaction gain (IG) to quantify the dependence of three variables $X$, $Y$, and $Z$ [27]:

$$ IG(X; Y; Z) = MI(X; Y; Z) - MI(X; Z) - MI(Y; Z) $$

(7)

where $X$ and $Y$ represent two genes, and $Z$ represents the class. To reflect the synergistic effect, each marginal effect of $X$ and $Y$ is subtracted from the mutual information. If $IG(X; Y; Z) > 0$, the ways in which the two genes $X$, $Y$ interact with each other can provide information that cannot be provided by either of them individually. Conversely, $IG(X; Y; Z) < 0$ when a redundancy exists between $X$ and $Y$.

To further detect the nonlinear direct association between gene $X$ and gene $Y$ when another variable $Z$ is given, conditional mutual information (CMI) was proposed. CMI is defined as the amount of information on gene $X$ and gene $Y$ when class $Z$ is given:

$$ CMI(X; Y | Z) = MI(X; Y) + H(Z | X) + H(Z | Y) - H(X, Y | Z) $$

(8)

where $p(x, y | z)$ is the joint probability distribution of $X$ and $Y$ with the condition $Z$, $p(z)$ is the probability of variable $Z$, and $p(x | z)$ and $p(y | z)$ are conditional marginal probability distributions.

Part mutual information (PMI) is proposed to quantify nonlinear direct dependencies between genes; this may overcome the false-positive problem of MI and CMI. PMI is defined as follows:

$$ PMI(X; Y | Z) = CMI(X; Y | Z) + \sum_{x \in X, z \in Z} p(x | z) \log \frac{p(x | z)}{p*(x | z)} $$

(9)

$$ + \sum_{y \in Y, z \in Z} p(y | z) \log \frac{p(y | z)}{p*(y | z)} $$

where $p*(x | z) = \sum_{y \in Y} p(x | y, z)p(y)$ and $p*(y | z) = \sum_{x \in X} p(y | x, z)p(x)$.

PMI can be used to quantify the nonlinear direct dependencies between genes $X$ and $Y$ when class $Z$ is given.

2) INTERACTION COEFFICIENT

From Eq (7), we can see that the introduction of gene $Y$ affects the dependence between gene $X$ and class $C$. A positive value of the interaction gain indicates an increasing amount of dependence, which means that the introduction of gene $Y$ has a positive influence in predicting the class $C$. Therefore, the weight of gene $Y$ should be increased. In contrast, a negative interaction gain indicates that introduction of the new gene will decrease the amount of dependence. Correspondingly, the weight of gene $Y$ should be decreased. Therefore, we introduce an interaction weight factor (IWF) [28] based on the interaction gain and use it to adjust the relationships between genes:

$$ IWF(X, Y) = 1 + \frac{IG(X; Y; Z)}{H(X) + H(Y)} $$

(10)

where $IG(X; Y; Z)$ is used to quantify the interaction gain between gene $X$ and gene $Y$ when class $Z$ is given, and $H(X)$ and $H(Y)$ quantify the average uncertainties of the two genes, respectively. However, IWF may not effectively quantify low associations.

To enhance the ability to detect weaker associations between two genes [29], IWF is modified to interaction coefficient (IC) in this paper:

$$ IC(X, Y) = 1 + \frac{IG(X; Y; Z)}{H(X, Y)} $$

(11)

$IC(X, Y)$ has the following properties:

Property 1. $0 \leq IC(X; Y) \leq 2$

Proof. By Eq (6), the following inequalities are obtained:

$$ 0 \leq MI(X; Y; Z) \leq H(X, Y) $$

(12)

$$ IG(X; Y; Z) \leq H(X, Y) $$

(13)

$$ MI(X; Z) + MI(Y; Z) \leq H(X, Y) $$

(14)

As Eq (7) and Eq (14), then

$$ MI(X; Y; Z) - MI(X; Z) - MI(Y; Z) \geq -H(X, Y) $$

(15)

$$ -H(X, Y) \leq IG(X; Y; Z) \leq H(X, Y) $$

(16)

Hence,

$$ 0 \leq 1 + \frac{IG(X; Y; Z)}{H(X, Y)} \leq 2 $$

(17)

according to the definition of interactive weight, we obtain $0 \leq IC(X, Y) \leq 2$.

Property 2. $0 \leq IC(X; Y) \leq 1$, when gene $X$ is redundant with gene $Y$.

Proof. If gene $X$ is redundant with gene $Y$, then

$$ IG(X; Y; Z) \leq 0 $$

(19)

As Eq (16), we have
\[
0 \leq I + \frac{IG(X;Y;Z)}{H(X,Y)} \leq 1 ,
\]
that is,
\[
0 \leq IC(X,Y) \leq 1
\]

**Property 3.** \( I \leq IC(X,Y) \leq 2 \), when gene \( X \) is interactive with gene \( Y \).

**Proof.** If gene \( X \) is interactive with gene \( Y \), that is \( IG(X;Y;Z) > 0 \) combining with Eq (16), we have
\[
0 \leq I + \frac{IG(X;Y;Z)}{H(X,Y)} \leq 2 ,
\]
which is equal to
\[
0 \leq IC(X,Y) \leq 2 .
\]

3) **INTERACTION PART MUTUAL INFORMATION**

Although mutual information has been widely used in constructing gene interaction networks, it ignores the effect of interaction between genes. To effectively quantify the interaction between genes, this paper proposes a new notion, interaction part mutual information (IPMI), based on IC and PMI:

\[
IPMI(X, Y | Z) = IC(X, Y) \ast (1 + PMI(X, Y | Z))
\]

IPMI can quantify nonlinear direct dependencies between genes. At the same time, IPMI regulates the weight of interactions between different genes based on the different conditions. If the information regarding the introduction of the new feature increases the amount of dependence between two genes when class \( Z \) is given, then the weight of PMI is increased by using IC; if the information regarding the introduction of the new feature decreases the amount of dependence when class \( Z \) is given, the weight of PMI is decreased.

In the genetic network, nodes and edges represent genes and the interactions between them, respectively. After sorting all the relevant values of IPMI, we performed curve fitting and then select the first inflection point as the weight of the edges in the network.

C. **IDENTIFICATION OF CHARACTERISTIC GENES AND PATHWAYS**

To detect characteristic genes, we defined the Topological Score (TS) of each node \( x \):

\[
TS_x = \frac{BC_x + CLC_x}{EC_x}
\]

where \( BC \), \( CLC \) and \( EC \) represent betweenness centrality, clustering coefficient, and eccentricity centrality, respectively. Among them, \( BC \) represents the sum of the proportions of the nodes appearing in the shortest path of the other nodes; it is defined as

\[
BC_x = \sum_{i:j \in V} \frac{N_{ij}}{N_{jj}}
\]

where \( N_{ij} \) represents the number of shortest paths from node \( i \) to node \( j \) and \( N_{jj} \) represents the number of paths through node \( v \). The Clustering Coefficient (CLC) is defined as

\[
CLC_v = \frac{2n}{k(k-1)}
\]

where \( k \) is the number of neighbors of node \( v \), and \( n \) is the number of edges connected between \( k \) neighbors of node \( v \). Furthermore, the eccentricity centrality of node \( v \) in the network \( G \) is defined as

\[
EC_v = \frac{1}{\max_{v \in G} (dist(v,u))}
\]

where \( u \) is a node in the network \( G \) and \( dist(v,u) \) is the largest geodesic distance between node \( v \) and other nodes.

The betweenness centrality of a node represents the role it plays in connecting with other nodes. The higher the value of BC, the more important the node is in maintaining network tightness; the clustering coefficient represents the density of the connection between nodes. The higher the value of CLC, the more closely the node is connected to its neighbors; the eccentricity centrality represents the largest distance from one node to others in the network. The smaller the value of EC, the more important the node is in the network. However, CLC and EC only reflect the local property, and BC only reflects the global property. Taking the advantages of these three topological properties into consideration, TS was proposed to detect characteristic genes effectively. TS combines the local property with the global property in the network simultaneously.

III. **RESULTS**

A. **DATASETS**

The Cancer Genome Atlas (TCGA) plan intends to create a map of all genome variations associated with human cancers using high-throughput sequencing technology [30]. In this paper, the experimental datasets were integrated gene expression data on three kinds of cancers downloaded from the TCGA database (https://tcgadata.nci.nih.gov/tcga/); the cancers included pancreatic cancer (PAAD), esophageal cancer (ESCA) and colorectal cancer (COAD). These datasets are summarized in Table I. A summary of the NIPMI analysis is shown in Fig. 1. In the experiment, Robust PCA (RPCA) was used to reduce dimensionality and to remove redundant information and noise from the integrated dataset. After the above preprocessing step, we obtained an integrated gene expression matrix containing 2532 rows and 621 columns.
B. NETWORK CONSTRUCTION AND ANALYSIS

1) COMPARISON OF IPMI WITH SEVERAL NETWORK CONSTRUCTION MEASURES

The IPMI measure was used to quantify the interactions between each pair of genes in the expression matrix that was obtained from the integrated dataset derived from the TCGA database. To make IPMI more effective and because the gene expression data are continuous, we approximated the continuous distribution by a finite discrete distribution. First, we used the four measures IPMI, PMI, CMI and IG to calculate correlation coefficient matrices between 2532 genes obtained by the RPCA feature selection method. Based on sorting the values of the IPMI coefficient matrix, we selected the value of IPMI that corresponds to the first inflection point (shown in Fig 2A) as the threshold for screening the network construction genes. In this way, 2199 gene nodes and 265,584 edges can be obtained. In the network, nodes represent genes, and edges represent interactions between two nodes. To further demonstrate the advantages of our method, we compared the network constructed by the IPMI measure to the networks constructed by the PMI, CMI, and IG measures. The three first inflection points based on the PMI, CMI, and IG coefficient matrixes, which were used as thresholds to construct the other three genetic interaction networks, are shown in Fig 2B, Fig 2C, and Fig 2D, respectively. Then, referring to the GeneCards database (http://www.GeneCardss.org), the number of network construction genes related to these three types of cancer that were obtained by the four measures was recorded and is shown in Table II. As seen from Table II, the network constructed by the IPMI measure contains the largest number of genes associated with these three cancers.

2) ANALYSES OF CHARACTERISTIC GENES

We analyzed the interactions among the genes in the above networks. The Topological Score of each node was used as an indicator of its contribution to the interactions among the genes in the network. The TS values of the nodes in the four networks were calculated. The larger a node’s TS value is, the more important the node is. Table III lists the number of top 10, top 30, top 50 and top 100 genes found in the GeneCards database. When multiple genes correspond to one value, we adopted the strategy of retaining the first value. From Table III, we can see that, overall, IPMI detected the highest number of genes simultaneously related to these three cancers.

![Diagram](Image)

**FIGURE 1.** Schematic diagram for network analysis of NIPMI. (1) shows three types of gene expression datasets for cancers. (2) is the integrated gene expression dataset, (3) represents the adjacency matrix that is obtained using the IPMI measure after selection by RPCA. (4) represents the gene interaction network. (5) shows characteristic genes detected by TS. (6) represents enriched pathway items.

**TABLE I**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Normal Samples</th>
<th>Tumor Samples</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAAD</td>
<td>176</td>
<td>4</td>
<td>20502</td>
</tr>
<tr>
<td>ESCA</td>
<td>183</td>
<td>9</td>
<td>20502</td>
</tr>
<tr>
<td>COAD</td>
<td>262</td>
<td>19</td>
<td>20502</td>
</tr>
<tr>
<td>Integrated</td>
<td>621</td>
<td>32</td>
<td>20502</td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>Measures</th>
<th>IPMI</th>
<th>PMI</th>
<th>CMI</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>722</td>
<td>709</td>
<td>387</td>
<td>588</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th></th>
<th>IPMI</th>
<th>PMI</th>
<th>CMI</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Top 30</td>
<td>13</td>
<td>6</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>Top 50</td>
<td>23</td>
<td>12</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>Top 100</td>
<td>44</td>
<td>27</td>
<td>41</td>
<td>42</td>
</tr>
</tbody>
</table>
FIGURE 2. Selection of inflection points. Fig 2A shows the inflection point (P) chosen for the IPMI measure. Fig 2B shows the inflection point (Q) chosen for the IPMI measure. Fig 2C shows the inflection point (R) chosen for the IPMI measure.

The genes belonging to the top 10 genes as calculated by TS in the network constructed by IPMI are listed in Table IV. The genes shown in bold have been recorded in the GeneCards database, and it has been confirmed that they are simultaneously involved in the development of pancreatic cancer, esophageal cancer and colorectal cancer. One of these genes, SFN, is a protein-coding gene associated with benign breast adenomyoeptithelioma and cocoon syndrome. SFN has been identified as a marker for screening pancreatic cancer samples by PCA [31] and has been used to obtain information regarding tumor classification and potential therapeutic response. The SFN gene product interferes with intracellular signaling pathways and cell cycle checkpoints and may indicate poor prognosis in human malignant tumors, including esophageal cancer [32]. KRT14 is a protein-coding gene that is associated with epidermolysis bullosa simplex, Dowling-Meara type epidermolysis bullosa simplex, Autosomal Recessive 1 and other conditions. In [33], it was shown that KRT14 is a diagnostic marker for potential esophageal squamous cell carcinoma; this finding may guide the future clinical treatment of esophageal cancer. In [34], it was demonstrated that gene polymorphism of MGST1, an environmental gene, may lead to colorectal cancer in people younger than 50 years of age. As a protein-coding gene, MMP7 has largely been studied in cancers, especially pancreatic cancer, which is the most common type of periampullary cancer. In [35], it was shown that MMP7 expression is associated with increased mortality in patients with pancreatic cancer. MMP7 may interact with MMP9 and MMP13 to affect early-stage esophageal cancer [36], and the co-expression of these three genes predicts poor outcome for esophageal cancer patients at a relatively early stage. MMP7, which is expressed in most tumor states, can act as a complement [37] in the identification of suspected locally advanced cancer. In addition, MMP7 expression in colorectal cancer is related to metastatic disease.

TABLE IV
NUMBERS OF CHARACTERISTIC GENES IN THE TOP 10 CAPTURED GENES BY FOUR MEASURES.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Count</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPMI</td>
<td>4</td>
<td>MIE1, STARD3, GJB6, SFN, KRT14, MGST1, MMP7, B3GNT5, LGALS7B, GPMB</td>
</tr>
<tr>
<td>PMI</td>
<td>2</td>
<td>PPP1R1B, MIE1, CX3C1J, COL11A1, ASPN, NAV2, APOD, WHSC1, KCNQ1, PLVAP</td>
</tr>
<tr>
<td>CMI</td>
<td>3</td>
<td>PGC, MT2A, FAM120A, PVRL1, PDIA2, AQP8, SYNGR2, EVPL, PRDX5, LSR</td>
</tr>
<tr>
<td>IG</td>
<td>4</td>
<td>TORA4, EPHA2, ACOX1, SLC22A23, KRT1, PMEP1, FSTL1, GSTM1, AQP5, GALNT6</td>
</tr>
</tbody>
</table>

Count: The number of genes that have been confirmed to be related to pancreatic cancer, esophageal cancer and colorectal cancer from top 10 genes obtained by Topological Score.

Similarly, genes belonging to the top 10 genes are recorded in Table IV; these genes were identified by TS in the other three networks constructed from PMI, CMI and IG, respectively. t-Darpp, a protein encoded by the PPP1R1B gene [38], is expressed in esophageal cancer, breast cancer, gastric cancer, and prostate cancers as well as in normal adult brain striatal cells. The COL11A1 gene [39] is overexpressed in pancreatic cancer and is a novel predictive biomarker for pancreatic cancer. COL11A1 [40] is the hub node in the PPI network and may serve as a target gene in esophageal cancer treatment. In [41], it was suggested that stromal expression of COL11A1 is related to malignancy of colorectal cancer. Overexpression of PGC [42] affects extracellular matrix degradation, which is one factor responsible for tumor progression and metastasis. In [43], it was suggested that the AQP8 gene is differentially expressed in colorectal tumors and normal colorectal epithelium, and the expression of this gene is a marker of normal proliferating colonic epithelial cells. In [44], it was
confirmed that the expression of AQP8 is associated with pancreatic tumor size, differentiation, lymph node metastasis, and TNM stage. APE1 inhibitors are used in the treatment of pancreatic cancer and other cancers [45]. The redox function of APE1 is responsible for driving the expression of mitochondrial genes such as PRDX5. EPHA2 [46], which is related to tumor differentiation and lymph node metastasis in esophageal cancer, is considered a potential target to prevent cancer cells from spreading into the lymphatic drainage. KRT11 is a protein-coding gene that is related to nonpidermolytic palmoplantar keratoderma and cyclic ichthyosis with epidermolytic hyperkeratosis. In [47], it was confirmed that some cancers, including colorectal, pancreatic, and esophageal cancers, may show homozygous deletions of GSTM1 and GSTT1. Overexpression of AQP5 has been shown to be related to multiple cancers such as pancreatic cancer and colon cancer [48]. From Table IV, we can see that the largest number of genes have been confirmed to be related to these cancers is found in the network constructed by the IPMI measure.

C. KEGG Pathway Enrichment Analysis
We utilized the KOBAS online analysis database (version 3.0) (http://kobas.cbi.pku.edu.cn/) to perform pathway enrichment analysis for the genes that were used to construct the network. The databases used in our pathway analysis included KEGG and Reactome; adjusted P-value < 0.05 was set as the cut-off threshold. The pathways in the four networks constructed using the IPMI, PMI, CMI, and IG measures were detected. The number of pathways found is shown in Table V. It is clear that the network constructed by the IPMI measure includes more pathways than the networks constructed by the other three measures.

<table>
<thead>
<tr>
<th>Measures</th>
<th>IPMI</th>
<th>PMI</th>
<th>CMI</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of pathways</td>
<td>292</td>
<td>289</td>
<td>257</td>
<td>286</td>
</tr>
</tbody>
</table>

To further verify the biological significance of the network constructed by the IPMI measure, we listed the details of the top 10 pathways according to their P-values in Fig 3. In the figure, the size of the nodes represents the number of genes that are enriched in individual pathways. The deeper the red color of the node, the smaller is the adjusted P-value of the pathway and the more important the pathway is. Among the pathways, focal adhesion (adjusted P-value=1.73E-20), metabolic pathways (adjusted P-value=1.44E-15), the PI3K-Akt signaling pathway (adjusted P-value=3.47E-155), and ECM-receptor interaction (adjusted P-value=1.10E-14) have been confirmed to be associated with pancreatic cancer, esophageal cancer and colorectal cancer. In [49], identified genes that are differentially expressed in pancreatic cancer were shown to be enriched in focal adhesion and other pathways. In [50], the PI3K-Akt signaling pathway was considered as a potential therapeutic target for the treatment of pancreatic cancer. The alteration of metabolic pathways in cancer has recently been recognized as an emerging biomarker of cancer [51, 52]. The ECM-receptor interaction pathway may exert direct or indirect control of cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis, and it has been confirmed to be related to lung cancer [53] and atrial fibrillation.

FIGURE 3. Top 10 significantly enriched pathway terms associated with network construction genes.

IV. CONCLUSION
Based on integrated gene expression data on multi-cancers, this paper proposed the NIPMI method, which attempts to identify the common characteristic genes that are associated with multicancer pathological processes. In this method, network construction genes are selected from integrated multi-cancers data using the RPCA feature selection method. Then, the proposed IPMI measure is used to construct a gene interaction network. We mined the information associated with nodes in the network and performed pathway enrichment analysis to extract cancer-related information from the network. Comparing the network construction method based on the IPMI measure with three other network construction measures, we find that there are more characteristic genes and more pathways in the network obtained using the IPMI measure.

The NIPMI method has several advantages. First, finding the characteristic genes by constructing a network and performing node mining not only considers the roles of single genes but also takes into account the interactions between genes. Second, use of the RPCA feature selection method filters out a large number of features, effectively reducing the difficulty of network construction. Third, the novel IPMI measure proposed in this paper can not only quantify the nonlinear direct dependence between genes but also reflect whether the information provided by a given gene is redundant or interactive; this enriches the information contained in the gene interaction network and facilitates the detection of key anomalous expression nodes throughout the network. However, this paper does not consider the higher-order interactions among genes, and further biological experiments are needed to verify the results of our experiment.
REFERENCES


QIAN DING received the B.S. degree in computer science and technology from Qufu Normal University, China, in 2018; the Master degree candidate in computer science and technology from Qufu Normal University, China. Her research interests include bioinformatics.

JUNLIANG SHANG received his B.S., M.S., and Ph.D. degrees from Xidian University, Xi’an, China, in 2007, 2010, and 2013, respectively. He is currently an associate professor in School of Information Science and Engineering, Qufu Normal University, Rizhao, China. His research interests focus on bioinformatics and big data mining.

YAN SUN received her B.S. degree from School of Computer Science, Shaanxi Normal University, Xi’an, China, in 2007, and received her M.S. degree from School of Economics and Management, Xi’an, China, in 2010. From 2010 to 2013, she was an intermediate editor in Shaanxi People’s Fine Arts Publishing House in Xi’an, China. She is currently a lecturer in School of Computer Science, Qufu Normal University, Rizhao, China. Her research interests focus on Intelligent Information Processing, Bioinformatics, and Massive Data Mining.

GUANGSHUAI LIU received his B.S. and M.S. degrees from Qufu Normal University in 2010 and 2013, respectively. He received his Ph.D. degree from Northeast Forestry University in 2016. He is currently a lecturer in the College of Life Science, Qufu Normal University, Qufu, China. His research interests focus on adaptive evolution.

FENG LI received the BSc and MSc degrees from Shandong University of Science and Technology, China, in 2009 and 2012, respectively. She received the doctor’s degree in computer application technology from Xidian University, China, in 2018. She is currently a lecturer in School of Information Science and Engineering, Qufu Normal University, Rizhao, China. Her current research interests include bioinformatics and data mining.

XIGUO YUAN received his B.S. and M.S. degree in computer applications from Wuhan University of Science & Technology in 2005 and 2008, respectively. He received his Ph.D. degree in computer applications from Xidian University in 2011. He is currently an associate professor at the School of Computer Science and Technology, Xidian University.

JINXING LIU received the B.S. degree in electronic information and electrical engineering from Shandong University, China, in 1997; 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 worked at Shenzhen graduate school, Harbin Institute of Technology as a postdoctoral research fellow. He is a professor at School of Information Science and Engineering, Qufu Normal University, Rizhao, China. His research interests include pattern recognition, machine learning, and bioinformatics.