Discovery of Relationships Between Long Non-coding RNAs and Genes in Human Diseases Based on Tensor Completion

Chen Peng¹, Liang Zou², De-Shuang Huang¹

¹Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai, China
²Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Corresponding author: DS Huang (dshuang@tongji.edu.cn)

This work is supported by National Natural Science Foundation of China [Grant Nos. 61702371, 61520106006, 61732012, 61772370, 31571364, 61532008, U1611265, 61672382, 61402334, 6147228061472280 and 61772357], and China Postdoctoral Science Foundation (Grant No. 2017M611619 and 2016M601646).

ABSTRACT

Thousands of long non-coding RNAs (lncRNAs) are encoded by mammalian genomes and play important roles in various biological processes including the regulation of gene transcription. Through these relationships, lncRNAs can participate in proliferation, differentiation and cytoprotective programs, which implies their critical roles in human diseases especially cancers. Therefore, there is an urgent need to study the relationships between lncRNAs and genes in human diseases, which will help uncover the mechanisms underlying disease progression. In this study, we explore the relationships between lncRNAs and genes in various diseases through a Tensor Completion based Approach (TCA). The results of performance evaluation suggest that TCA can obtain a significantly better performance than the baseline method. Moreover, top ranked relationships with highest average scores in all diseases corroborate the effectiveness of TCA and the reliability of the predicted results. Case study of hepatocellular carcinoma (HCC) indicate that the elements of the top ranked relationships may be functionally implicated in HCC. Furthermore, three lncRNAs (HULC, MALAT1 and BANCR) and gene HOXB7 are found to be important in HCC, which are consistent with the newest reports and existing literatures.

INDEX TERMS

Diseases, lncRNAs, tensor, cancers, genes

I. INTRODUCTION

Long non-coding RNAs (lncRNAs) are long, polyadenylated RNAs that do not code for proteins. Based on RNA-seq data, large-scale analyses of chromatin-state maps [1]-[2], full-length complementary DNA sequence [3]-[5] and other analyses have identified a lot of lncRNAs [6]. These non-coding transcripts are widely involved in various cellular processes including imprinting control, cell differentiation, human diseases, tumorigenesis and immune responses [7]-[11], which indicate the widespread functionality and complex metabolism of lncRNAs. Besides, more and more evidence suggests that by means of influencing transcriptional protein translation, mRNA splicing, complexes targeting and chromatin modification [12]-[13], lncRNAs play important roles in regulating gene expressions at transcriptional, post-transcriptional and epigenetic levels [14], and are increasingly considered as functional regulatory components of gene regulation [15]. These relationships between lncRNAs and genes are initially found by the researches of mammalian X-chromosome inactivation (XCI) [16], and through this way lncRNAs can participate in proliferation, differentiation and cytoprotective programs, which implies their potential roles in human disease especially cancers. It is reported that the alterations of lncRNAs are related to transcriptional silencing in many types of cancer and may result in inappropriate de-repression of gene expression. Therefore, there is an urgent need to study the relationships between lncRNAs and genes in human diseases, which will help uncover the mechanisms underlying disease progression. Recently, a growing number of researches utilize either low-throughput or high-throughput experimental technologies to explore the relationships between lncRNAs and genes in human diseases [14]. The former technologies include western blot and quantitative reverse transcriptase-polymerase chain reaction while the latter technologies...
include RNA-seq and microarray [14]. According to the reports that lncRNAs are associated with malignancies via aberrant expression patterns in most cases [17], these studies focus on identifying genes (genome-wide or individual) that are differentially expressed after overexpressing or knocking down a lncRNA under the condition of various diseases and these genes are then regarded as targets of the lncRNA. Many lncRNAs and their relationships with genes are already linked to the processes of different cancers [17] such as HOTAIR (implicated in gastrointestinal, pancreatic, hepatic and colorectal cancers [18]-[21]), MALAT1 (implicated in non-small cell lung cancer and hepatoblastoma [22]-[23]), TUG1 (implicated in bladder cancer [24]), LSINCT5 (implicated in ovarian and breast cancer [25]), UCA1 (implicated in squamous carcinoma [26]), and so on. In addition, overexpression of lncRNA ANRIL is found in prostate cancer and the loss of its expression correlates with a reduction of cellular lifespan [17]. Meanwhile, the expressions of tumor suppressor genes INK4A and INK4B, which are respectively encoded by CDKN2A and CDKN2B, are increased after knockdown of ANRIL [27]. Existing researches based on low-throughput or high-throughput technologies make great contribution to the studies of relationships between lncRNAs and genes in human diseases especially cancers. Nevertheless, biological experiments [25], [28] are expensive, time-consuming and hard to operate, which restrict the rapid development of lncRNA-related exploration. Moreover, bioinformatics tools based on high-throughput data are inconvenient to some extent due to the requirement of large-scale data downloaded from well-established databases [14], [29]-[34]. Most importantly, these researches generally focus on a few lncRNAs and genes in one certain disease, which cannot provide an overall analysis of the relationships between lncRNAs and genes in different diseases.

In this study, we explore the relationships between lncRNAs and genes in human diseases through a Tensor Completion based Approach (TCA). The novelty of this approach lies in that the relationships between lncRNAs, genes and diseases are simultaneously analyzed based on tensor, and then novel relationships are directly obtained. Furthermore, large-scale high-throughput data is not required in the approach, which improves its practicability and convenience. The experiments of performance evaluation suggest a significantly better performance of TCA than the baseline method. Moreover, the statistical analysis of unique relationships in each disease shows that hepatocellular carcinoma (HCC) has the largest number of relationships that different from those normal ones. Subsequent functional enrichment analysis and network analysis indicate that the elements of top ranked relationships may be functionally implicated in HCC. Besides, three lncRNAs (HULC, MALAT1 and BANCR) and gene HOXB7 are found to be important in HCC, which are consistent with the newest reports and existing literatures.

II. MATERIALS AND METHODS

A. KNOWN RELATIONSHIPS BETWEEN LNCRNAS AND GENES IN HUMAN DISEASES

We collect known experimentally verified relationships between lncRNAs and genes in human diseases from a public database LncRNA2Target (http://www.lncrna2target.org), in which genes that are differentially expressed after knockdown or overexpression of a certain lncRNA are regarded as the targets of this lncRNA. Specifically, two data sets that store relationships obtained in different ways are downloaded from this website. One is derived from low-throughput experimental technologies such as western blot or quantitative reverse transcriptase-polymerase chain reaction while the other is from high-throughput experiment technologies including RNA-seq and microarray [14]. Based on the data set from low-throughput technologies, we can obtain abundant information about the relationships, such as lncRNA symbol, genomic location, corresponding target gene symbol, disease state, reference information and cell line in which the experiments are performed [14], etc. Meanwhile, the data set from high-throughput technologies also provides the basic information of lncRNAs and target genes, as well as statistic parameters utilized in the experiments including the value of log2FC, p-value and adjusted p-value. By integrating above two data sets, totally 852 known experimentally verified relationships between lncRNAs and genes in human diseases are obtained. For convenience, the relationships derived from normal tissues are thereafter named normal interactions.

B. TENSOR MODELING OF RELATIONSHIPS BETWEEN LNCRNAS AND GENES IN HUMAN DISEASES

The discovery of relationships between lncRNAs and genes in human diseases is an issue of prediction of ternary relations between three factors, i.e., lncRNAs, genes and diseases. Based on known relationships in 32 human diseases and normal tissues, we make use of a three-way tensor $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$ to model the relationships between these three factors, in which $I$, $J$ and $K$ represent the numbers of lncRNAs, target genes and diseases, respectively. The triples in $\mathbf{X}$ are the known relationships, i.e., positive observations in the past (Fig. 1). Since that lots of entries in $\mathbf{X}$ are unknown due to the insufficiency of experimental data, our goal is to recover these missing information and complete the tensor.
C. TCA FOR DISCOVERY OF RELATIONSHIPS BETWEEN lncRNAs AND GENES IN HUMAN DISEASES

In this study, the relationships between lncRNAs and genes in human diseases are predicted by TCA, in which by using tensor model the underlying factors in each dimension can be extracted on the premise of not damaging the multi-way nature of data. Here a powerful technique, i.e., CANDECOMP/PARAFAC (CP) tensor decomposition [35]-[38], is utilized. For the three-way tensor $\mathcal{X}$ with size $I \times J \times K$, we assume that its rank is $R$ and the CP decomposition is defined as tensor factorization, i.e., if data is perfect, tensor $\mathcal{X}$ can be modeled by three factor matrices $A \in \mathbb{R}^{I \times R}$, $B \in \mathbb{R}^{J \times R}$ and $C \in \mathbb{R}^{K \times R}$ as follows:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \quad \text{for all } i=1,\ldots,I, j=1,\ldots,J, k=1,\ldots,K \quad (1)$$

where $x_{ijk}$, $a_{ir}$, $b_{jr}$ and $c_{kr}$ are the elements of tensor $\mathcal{X}$, matrix $A$, matrix $B$ and matrix $C$, respectively. Actually, in most cases the data occurs in the presence of noise and true $\mathcal{X}$ is not observable, so equation (1) is not applicable under this circumstance. Besides, when the data is incomplete such as $\mathcal{X}$, in which lots of entries are unknown, the implementation of CP decomposition turns into an optimization problem with the goal of minimizing an error function. In order to model only known entries and ignore missing data, we take advantage of a weighted version of error function, which can be calculated by the following equation [39]:

$$f_w(A,B,C) = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} w_{ijk} \left( x_{ijk} - \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \right)^2 \quad (2)$$

Meanwhile, the objective function is defined as [40]:

$$\min_{\mathcal{X}} \frac{1}{2} \| \mathcal{W} (\mathcal{X} - \hat{\mathcal{X}}) \|_F^2 \quad (3)$$

where $\hat{\mathcal{X}}$ is the tensor that best approximates $\mathcal{X}$ with $R$ components and $\mathcal{W}$ is a nonnegative weigh tensor with the same size as $\mathcal{X}$. Besides, the symbol “ $\circ$ ” indicates the operation of outer product while $a_i$, $b_j$ and $c_k$ denote the $i$-th column of factor matrices $A$, $B$ and $C$ [41]-[43], respectively. The parameter $w_{ijk}$ in equation (2) is the element of $\mathcal{W}$ and calculated as follows [39]:

$$w_{ijk} = \begin{cases} 1 & \text{ if } x_{ijk} \text{ is known,} \\ 0 & \text{ if } x_{ijk} \text{ is unknown,} \end{cases} \quad \text{for all } i=1,\ldots,I, j=1,\ldots,J, k=1,\ldots,K \quad (4)$$

This optimization problem with a weighted least squares objective function can be solved by any gradient-based optimization method when the gradient and function is available [39]. Then the factors obtained are used to reconstruct the tensor and predict the relationships between lncRNAs and genes in human diseases.

D. THE PERFORMANCE EVALUATION

To evaluate the performance of TCA in discovering relationships between lncRNAs and genes in human diseases, we utilize n-fold cross-validation in the test process. Firstly, all known relationships are evenly divided into $n$ parts. Then in each round, the known information for one part of relationships are removed and these relationships are taken as test data that predicted based on all other n-1 parts of relationships. When the value of $n$ equals to the number of known relationships, the n-fold cross-validation is also called leave-one-out cross-validation (LOOCV). Known and unknown relationships are regarded as golden standard positive (GSP) and golden standard negative (GSN) data, respectively. In this study, the top $k$ ranked results according to the scores predicted by TCA are defined as relationships between lncRNAs and genes in the corresponding disease. The intersections of top $k$ ranked results with GSP data are considered as true positive (TP) data and the remaining part of GSP are referred to as false negatives (FN) data. Similarly, the intersections of top $k$ ranked results with GSN data and the remaining part of GSN are considered as false positive (FP) data and true negative (TN) data, respectively. Based on these information, the values of specificity ($Sp$) and sensitivity ($Sn$) can be calculated as follows:

$$Sp = \frac{TN}{TN + FP} \quad Sn = \frac{TP}{TP + FN} \quad (5)$$

We take advantage of a commonly used measurement Receiver Operating Characteristic curve (ROC curve) to evaluate the performance of TCA. The x axis and y axis respectively denote 1-$Sp$ and $Sn$, and larger area under the curve (AUC) indicates higher accuracy. In addition, the values of precision and recall are also computed with the rank threshold varying from 1000 to 2000. The detailed definitions are shown in the following equation:

$$\text{precision} = \frac{TP}{TP + FP} \quad \text{recall} = \frac{TP}{TP + FN} \quad (6)$$

III. RESULTS

A. PERFORMANCE OF TCA

Due to the fact that there are few computational studies about the relationships between lncRNAs and genes in different diseases, we cannot compare the performance of TCA with other methods. However, in order to facilitate a fair performance assessment, we utilize a baseline method, in which the tensor used in TCA is randomly permuted 100 times. For each time, the process of tensor completion is repeated using the permuted tensor and tested with the same assessment procedure. Then the performance of baseline method is obtained by averaging the predicted results of all permutation tests (Fig. 2). As shown in Fig. 2(a), the ROC curves of TCA are all obviously above those of baseline based on 100 experiments. Specifically, the average AUC value of TCA is 91.3%, which is 40.0% higher than that of
FIGURE 2. The results of performance evaluation of TCA and baseline method. (a) ROC curves of 100 experiments. The gray curves represent the ROC curves of 100 experiments. The red and blue curves represent average ROC curves of TCA and baseline method, respectively. (b) Average AUC values and Sn values at three stringent levels of Sp of TCA and baseline method.

baseline. Besides, the Sn values of TCA at three stringent levels of Sp significantly exceed those of baseline as well (Fig. 2(b)). When Sp is high, i.e., 99.0%, the Sn values of TCA and baseline are 74.8% and 0.6%, respectively, which indicate the excellent performance of TCA under the condition of high specificity. The Sn value of baseline rises to 4.1% when Sp changes to 95.0%, which is 76.7% lower than that of TCA. When Sp drops to 90.0%, the Sn value of baseline increases by more than 100%, but is still far lower than that of TCA. These results suggest the effectiveness of TCA in predicting relationships between lncRNAs and genes in human diseases.

We also perform robustness analysis by conducting experiments with 5-fold, 10-fold, 50-fold, 100-fold cross validation and LOOCV. As shown in Fig. 3, the results of all cross-validation tests are qualitatively similar. Specifically, their AUC values are approximately 90% with a fluctuation range less than 1.5%, among which LOOCV obtains a slightly higher AUC value than other four experiments (Fig. 3(a)). Besides, the Sn values of all tests are larger than 70.0% when Sp is 99.0% and even exceed 80.0% when Sp is 90.0%. Moreover, the precision-recall curves of all cross-validation tests have similar trends and stay close to each other (Fig. 3(b)). Although the performance of 5-fold, 10-fold and 100-fold cross-validation experiments are not as excellent as that of 50-fold cross-validation and LOOCV, the precisions of them are always higher than 55.0% as rank threshold varying from 1000 to 2000. With regard to the result of LOOCV test, a precision of 68.1% and a recall of 58.0% can be achieved for the rank threshold 1000. When the threshold increases to 2000, TCA can still obtain a precision of 70.9% and a recall of 30.2%. By the way, we will use the results of LOOCV for further analysis.

B. PERFORMANCE EVALUATION WITH DIFFERENT NUMBERS OF KNOWN ENTRIES IN THE TENSOR

Due to the fact that TCA can efficiently recover the missing information from sparse data, e.g., the proportion of known entries is 0.2% in this study [39], we evaluate its performance
in more extreme cases. Here we gradually reduce the number of known entries (from 1% to 99%) in the original tensor and try to recover the missing information in these more sparse data. The experiment on each dataset is repeated for 10 times, and the boxplot of AUC values and the variation of mean AUC are shown in Fig. 4. From Fig. 4, we can see that mean AUC decreases as the number of known entries reduces on the whole. Initially, mean AUC keeps high (close to 0.9) with the reduction of the number of known entries less than 10%. Then the declining trend of mean AUC becomes obvious and it drops to 0.8 when 40% of known entries are removed from the original data. When we retain less than 20% of known entries, mean AUC decrease to about 0.5, which even approximates the result of random experiment. In a word, if the tensor has few known entries, e.g., the slice for each cancer has only one known entry, it would not be possible to faithfully recover the missing information in these extremely sparse data by TCA. The above results may also serve as guideline for the usage of TCA.

C. IDENTIFICATION OF NOVEL RELATIONSHIPS BETWEEN LncRNAs AND GENES IN HUMAN DISEASES

Based on the predicted results of TCA, we obtain the scores of the relationships between lncRNAs and genes in human diseases, which are used to measure the reliabilities of these relationships. To be fair, the scores of those known relationships are not involved in subsequent analysis. Firstly, the average score of each relationship in different diseases is calculated and ranked in descending order. Then we extract the top 100 ranked relationships and list their names as well as normalized scores in Table I. Due to that higher average scores may imply more important roles in all 32 diseases. As shown in Table I, the relationship between lncRNA MALAT1 and gene MMP9 achieves the highest score. Interestingly, there are already several literatures [44]-[46] reporting this regulation in human cancers. For example, Wu et al find that the expression of MMP9 is significantly down-regulated after MALAT1 knockdown in gallbladder cancer [44]. Similarly, Jiao et al also detect decreased expression of MMP9 in pancreatic cancer after suppressing the expression of MALAT1 [45]. Besides, in glioma cells MALAT1 acts as tumor suppressor by down-regulating MMP2 and inactivating ERK/MAPK signaling [46]. These discoveries corroborate the effectiveness of TCA and are evidence for the reliability of the predicted results.

In this study, predicted relationships with scores larger than a certain threshold are regarded as novel predicted relationships. As the threshold varying from 0.1 to 0.9, we extract unique relationships in each disease that are different from those normal interactions. Meanwhile, we perform statistical analysis by counting the numbers of relationships under different thresholds. The distribution of the numbers in different diseases are further exhibited by box plot. As shown in Fig. 5, obviously larger numbers of unique relationships are obtained in HCC, lung cancer and non-small cell lung cancer (NSCLC) than other 29 diseases, which may suggest that relationships between lncRNAs and genes play more important roles in the progression of these three cancers. Moreover, the number of relationships of HCC is largest with a medium value of 160 and 6 relationships can still be achieved when the threshold of score is set to 0.8. This phenomenon indicates that plenty of relationships between lncRNAs and genes are altered in HCC based on the comparison with normal interactions, and this kind of relationships may have a significant influence on HCC.
D. CASE STUDY: RELATIONSHIPS BETWEEN LncRNAs AND GENES IN HCC

The top 100 ranked relationships based on the mean scores in human diseases

<table>
<thead>
<tr>
<th>LncRNA</th>
<th>Gene</th>
<th>Normalized score</th>
<th>LncRNA</th>
<th>Gene</th>
<th>Normalized score</th>
<th>LncRNA</th>
<th>Gene</th>
<th>Normalized score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALAT1</td>
<td>MMP9</td>
<td>1.000</td>
<td>NEAT1</td>
<td>LPAR1</td>
<td>0.819</td>
<td>MALAT1</td>
<td>CASP8</td>
<td>0.785</td>
</tr>
<tr>
<td>NEAT1</td>
<td>MMP9</td>
<td>0.939</td>
<td>ANCR</td>
<td>HOTAIR</td>
<td>0.819</td>
<td>BANCRI</td>
<td>CDX2</td>
<td>0.785</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>PGK1</td>
<td>0.934</td>
<td>HOTAIR</td>
<td>GLUT5C1</td>
<td>0.817</td>
<td>HOTAIR</td>
<td>ACTR2</td>
<td>0.784</td>
</tr>
<tr>
<td>MALAT1</td>
<td>HOXA7</td>
<td>0.921</td>
<td>BANCRI</td>
<td>ABCA1</td>
<td>0.814</td>
<td>NEAT1</td>
<td>AKT3</td>
<td>0.784</td>
</tr>
<tr>
<td>ANCR</td>
<td>MMP9</td>
<td>0.913</td>
<td>HOTAIR</td>
<td>HOXA10</td>
<td>0.814</td>
<td>NEAT1</td>
<td>MBP</td>
<td>0.784</td>
</tr>
<tr>
<td>NEAT1</td>
<td>MYC</td>
<td>0.908</td>
<td>TUG1</td>
<td>PDGFB</td>
<td>0.814</td>
<td>NEAT1</td>
<td>CDCP1</td>
<td>0.784</td>
</tr>
<tr>
<td>TUG1</td>
<td>MMP9</td>
<td>0.899</td>
<td>ANCR</td>
<td>HOTAIR</td>
<td>0.813</td>
<td>HOTAIR</td>
<td>H19</td>
<td>0.782</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>MYC</td>
<td>0.896</td>
<td>MALAT1</td>
<td>TP53</td>
<td>0.813</td>
<td>HOTAIR</td>
<td>BMPER</td>
<td>0.781</td>
</tr>
<tr>
<td>MALAT1</td>
<td>BACE1</td>
<td>0.889</td>
<td>NEAT1</td>
<td>CDK2B</td>
<td>0.812</td>
<td>BANCRI</td>
<td>HMMR</td>
<td>0.781</td>
</tr>
<tr>
<td>ANCR</td>
<td>MYC</td>
<td>0.888</td>
<td>TUG1</td>
<td>HOXA7</td>
<td>0.810</td>
<td>TUG1</td>
<td>LPHN2</td>
<td>0.780</td>
</tr>
<tr>
<td>NEAT1</td>
<td>HOXB7</td>
<td>0.881</td>
<td>BANCRI</td>
<td>STC1</td>
<td>0.809</td>
<td>ANCR</td>
<td>TP53</td>
<td>0.780</td>
</tr>
<tr>
<td>NEAT1</td>
<td>CASP8</td>
<td>0.879</td>
<td>BANCRI</td>
<td>COL6A1</td>
<td>0.808</td>
<td>NEAT1</td>
<td>ARNT2</td>
<td>0.780</td>
</tr>
<tr>
<td>NEAT1</td>
<td>TP53</td>
<td>0.877</td>
<td>NEAT1</td>
<td>MMP2</td>
<td>0.806</td>
<td>BANCRI</td>
<td>GPC6</td>
<td>0.779</td>
</tr>
<tr>
<td>NEAT1</td>
<td>PDGFB</td>
<td>0.862</td>
<td>ANCR</td>
<td>PDGFB</td>
<td>0.805</td>
<td>BANCRI</td>
<td>MYC</td>
<td>0.778</td>
</tr>
<tr>
<td>BANCRI</td>
<td>LPHN2</td>
<td>0.860</td>
<td>HOTAIR</td>
<td>MAPK9</td>
<td>0.803</td>
<td>NEAT1</td>
<td>MAPK9</td>
<td>0.777</td>
</tr>
<tr>
<td>BANCRI</td>
<td>LPAR1</td>
<td>0.858</td>
<td>TUG1</td>
<td>TP53</td>
<td>0.802</td>
<td>HOTAIR</td>
<td>CPM</td>
<td>0.776</td>
</tr>
<tr>
<td>TUG1</td>
<td>CASP8</td>
<td>0.858</td>
<td>BANCRI</td>
<td>QKI</td>
<td>0.802</td>
<td>NEAT1</td>
<td>LPHN2</td>
<td>0.776</td>
</tr>
<tr>
<td>TUG1</td>
<td>MYC</td>
<td>0.857</td>
<td>MALAT1</td>
<td>FAS</td>
<td>0.802</td>
<td>TUG1</td>
<td>ROBO1</td>
<td>0.776</td>
</tr>
<tr>
<td>MALAT1</td>
<td>HOXB7</td>
<td>0.850</td>
<td>NEAT1</td>
<td>BACE1</td>
<td>0.801</td>
<td>BANCRI</td>
<td>HOXA10</td>
<td>0.776</td>
</tr>
<tr>
<td>BANCRI</td>
<td>MCAM</td>
<td>0.846</td>
<td>NEAT1</td>
<td>FZD5</td>
<td>0.799</td>
<td>NEAT1</td>
<td>IWS1</td>
<td>0.775</td>
</tr>
<tr>
<td>NEAT1</td>
<td>FGF1</td>
<td>0.840</td>
<td>HOTAIR</td>
<td>PRKA1</td>
<td>0.798</td>
<td>BANCRI</td>
<td>NNMT</td>
<td>0.774</td>
</tr>
<tr>
<td>BANCRI</td>
<td>ROBO1</td>
<td>0.839</td>
<td>HOTAIR</td>
<td>PDK1</td>
<td>0.798</td>
<td>HOTAIR</td>
<td>ANKFY1</td>
<td>0.774</td>
</tr>
<tr>
<td>BANCRI</td>
<td>CDCP1</td>
<td>0.838</td>
<td>HOTAIR</td>
<td>MBP</td>
<td>0.795</td>
<td>NEAT1</td>
<td>PRKAA1</td>
<td>0.772</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>HOXA7</td>
<td>0.836</td>
<td>BANCRI</td>
<td>CSF1</td>
<td>0.795</td>
<td>TUG1</td>
<td>CDK2B</td>
<td>0.772</td>
</tr>
<tr>
<td>BANCRI</td>
<td>MDM2</td>
<td>0.835</td>
<td>BANCRI</td>
<td>CASP8</td>
<td>0.794</td>
<td>BANCRI</td>
<td>BMPER</td>
<td>0.772</td>
</tr>
<tr>
<td>TUG1</td>
<td>LPAR1</td>
<td>0.833</td>
<td>HOTAIR</td>
<td>NDK2</td>
<td>0.792</td>
<td>MALAT1</td>
<td>HOXA10</td>
<td>0.772</td>
</tr>
<tr>
<td>MALAT1</td>
<td>PDGFB</td>
<td>0.829</td>
<td>TUG1</td>
<td>FGF1</td>
<td>0.792</td>
<td>BANCRI</td>
<td>BAX</td>
<td>0.770</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>MMP2</td>
<td>0.826</td>
<td>MALAT1</td>
<td>MDM2</td>
<td>0.790</td>
<td>BANCRI</td>
<td>CASP3</td>
<td>0.770</td>
</tr>
<tr>
<td>MALAT1</td>
<td>AKT3</td>
<td>0.826</td>
<td>ANCR</td>
<td>LPAR1</td>
<td>0.789</td>
<td>BANCRI</td>
<td>AKT3</td>
<td>0.768</td>
</tr>
<tr>
<td>BANCRI</td>
<td>HOXA7</td>
<td>0.825</td>
<td>BANCRI</td>
<td>CPM</td>
<td>0.788</td>
<td>ANCR</td>
<td>FGF1</td>
<td>0.768</td>
</tr>
<tr>
<td>BANCRI</td>
<td>CXCL5</td>
<td>0.824</td>
<td>BANCRI</td>
<td>PDGFB</td>
<td>0.788</td>
<td>ANCR</td>
<td>MMP2</td>
<td>0.767</td>
</tr>
<tr>
<td>NEAT1</td>
<td>HOXA7</td>
<td>0.823</td>
<td>TUG1</td>
<td>CDCP1</td>
<td>0.787</td>
<td>BANCRI</td>
<td>FZD5</td>
<td>0.766</td>
</tr>
<tr>
<td>ANCR</td>
<td>CASP8</td>
<td>0.822</td>
<td>MALAT1</td>
<td>FZD5</td>
<td>0.786</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOTAIR</td>
<td>BDNF</td>
<td>0.821</td>
<td>TUG1</td>
<td>CPM</td>
<td>0.785</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 5.** The distribution of the numbers of relationships under different thresholds in different diseases.
Since relationships between lncRNAs and genes may play important roles in HCC, we further extract the top 100 ranked relationships in HCC for subsequent analysis. Here functional enrichment analysis of the elements of these 100 relationships, which consist of 11 lncRNAs and 65 target genes, is performed by using DAVID. From the categories listed in Table II, we interestingly find that there are two terms directly related to HCC: “hsa05161: Hepatitis B” ($p$-value $= 3.3 \times 10^{-10}$) and “hsa04932: Non-alcoholic fatty liver disease (NAFLD)” ($p$-value $= 2.1 \times 10^{-5}$), which may corroborate the reliability of the predicted results in this study. Besides, several common biological activities and functions in human cancers appear in the results, e.g., “GO: 0008284—positive regulation of cell proliferation” ($p$-value $= 3.3 \times 10^{-8}$), “hsa05202: Transcriptional misregulation in cancer” ($p$-value $= 3.4 \times 10^{-7}$), “hsa05203: Viral carcinogenesis” ($p$-value $= 1.8 \times 10^{-5}$), “GO: 0042981—regulation of apoptotic process” ($p$-value $= 7.4 \times 10^{-5}$), “Tumor suppressor” ($p$-value $= 1.7 \times 10^{-8}$), “Disease mutation” ($p$-value $= 1.7 \times 10^{-8}$) and “Proto-oncogene” ($p$-value $= 3.0 \times 10^{-2}$). Moreover, as many as 13 types of cancers and 15 types of typical cancer-related pathways are listed in Table II, among which P53 [47], MAPK [48], PI3K-AKT [49], TNF [50], HIF-1 [51], WNT [52], NOTCH [53], TGF-beta [54], RAS [55] and MTOR [49] signaling pathways are related to HCC according to existing reports in several literatures [47]-[55]. All these results indicate the importance of the elements of top 100 ranked relationships between lncRNAs and genes in HCC. In addition, although some lncRNAs and genes are not directly associated with HCC, the functions in other cancers and cancer-related pathways may imply their potential roles in the progression of HCC.

We further build the network of top 100 ranked relationships in HCC, which consists of 11 lncRNAs and 65 target genes. As shown in Fig. 6, HULC, MALAT1 and BANCR are three lncRNAs that have the largest numbers of target genes. As shown in Fig. 6, HULC, MALAT1 and BANCR are three lncRNAs that have the largest numbers of target genes.
and the work of Li et al. also shows that the plasma level of HULC achieves a good accuracy in predicting diagnosis of HCC [57]. Similarly, Lai et al. suggest that MALAT1 could act as novel biomarker and therapeutic target for the prediction of tumor recurrence of HCC [58]. Meanwhile, the study in [59] indicates the associations between plasma MALAT1 level and liver damage, and point out the clinical utility of MALAT1 in predicting the development of HCC. With regard to BANCR, both Zhou [60] et al. and Qin [61] et al. point out its potential significance in the diagnosis and prognosis of HCC based on the observed contribution in initiation and progression of this cancer.

To analyze above three lncRNAs in detail, we further extract the predicted relationships of them (Table. III) and the subnetwork is constructed (Fig. 7), in which the target gene numbers of HULC, MALAT1 and BANCR are 43, 20 and 13, respectively. In Table. III, target genes are ranked by their scores in the predicted results, among which HOXB7 has the highest score in both gene lists of HULC and BANCR. Meanwhile, HOXB7 is the second ranked target gene in the list of MALAT1. This phenomenon indicates the critical role of HOXB7 in relationships between lncRNAs and genes, which may also imply its potential function in HCC. According to the records in DisGeNET (v5.0) (http://www.disgenet.org/web/DisGeNET), which is the largest public database collecting the associations between genes and human diseases, HOXB7 is related to as many as 46 cancers. By consulting literatures, there are two latest reports [62]-[63] showing that gene HOXB7 may be important in HCC. Specifically, in the study of [62], cell proliferation, migration and invasion of HCC can be promoted by HOXB7 via the activation of bFGF-induced MAPK/ERK pathway and the authors regard HOXB7 as promising prognostic factor as well as therapeutic target for HCC. In another latest literature [63], Huan et al. detect significant correlations between poor prognosis of HCC and overexpression of HOXB7. Moreover, malignant progression of HCC is accelerated by HOXB7 through promoting epithelial-mesenchymal transition and stemness of hepatoma cells [63]. All these phenomenons confirm the superiority of TCA in identifying relationships between lncRNAs and genes in human diseases.

IV. DISCUSSION AND CONCLUSION
A growing number of publications suggest that lncRNAs play nonnegligible roles in regulating gene expressions positively or negatively at transcriptional, post-transcriptional and epigenetic levels [64]-[67], and these relationships may occur during a wide range of tumorigenesis [68]. However, the potential biological functions of most lncRNAs and the regulatory mechanisms between them and genes especially under the condition of various diseases are still obscure and poorly understood. In this study, we address this issue by an approach based on tensor completion, in which the relationships between lncRNAs, genes and diseases are modeled in a recommender system with the goal of suggesting a lncRNA the genes it might target in a specific cancer. Then the problem of discovering relationships

![FIGURE 6. The network of top ranked 100 relationships in HCC.](image)

![FIGURE 7. The subnetwork consists of relationships of HULC, MALAT1 and BANCR.](image)

<table>
<thead>
<tr>
<th>LncRNA</th>
<th>Target genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HULC</td>
<td>HOXB7, HOXA10, PTEN, MAPK1, MYC, BAX, FAS, HOXA7, HMG2A, MDM2, HOXA1, CASP8, CDKN2B, VAMP3, MEF2C, SNAI2, BDNF, ATM, CASP9, FN1, QKI, ANKRY1, CASP3, NEAT1, ADAM12, CCND2, MAML1, AKT3, MMP3, FANK1, MMP9, FDZ5, AR, ARNT2, PDGF, MBP, FSCN1, CCNB1, TGFBI, COL6A1, VEGFA, FGFI, CXCL5</td>
</tr>
<tr>
<td>MALAT1</td>
<td>MKNK2, HOXB7, MYC, ANKRD11, SDC4, RPRD1A, TBC8, MORC2, IWS1, NNT, NDUFC2, FANK1, SNAI2, SNAPP2, ADAM12, NF195, LTBRI, BAX, CASP8, GLTSCR2</td>
</tr>
<tr>
<td>BANCR</td>
<td>HOXB7, GLTSCR2, ANKFY1, MKNK2, ATM, AR, MDM2, AKT3, SDC4, CASP7, P16, FANK1, FZD5</td>
</tr>
</tbody>
</table>

![TABLE III. Relationships of lncRNAs HULC, MALAT1 and BANCR](table)
between lncRNAs and genes in human diseases is transformed into the process of recovering missing information in the tensor model, i.e., tensor completion. It is the first time that tensor completion is used in modeling the relationships between lncRNAs and genes in human diseases. This approach takes advantage of known relationships between lncRNAs and genes in different diseases and avoids utilizing large-scale high-throughput data, which is inconvenient to obtain and preprocess. The results show that TCA can efficiently infer reliable relationships between lncRNAs and genes in human diseases with the evidence from newest reports and existing literatures. Besides, important lncRNAs and genes are extracted from the results according to functional enrichment analysis and network analysis, which can also help uncover the mechanisms underlying cancer progression.

Besides the discoveries discussed in Section “Results”, some other findings obtained in our work are also consistent with existing studies. For example, the relationships between lncRNA NEAT1 and gene MMP9 achieves the second largest average score for 32 types of diseases, which is in accordance with the conclusions in the studies of [69]-[70] that silencing of NEAT1 expression and overexpression of MMP1 respectively lead to the suppression and increase of MMP9 protein expression in non-small cell lung cancer cells. Moreover, in this study lncRNA BANC is found to have the largest number of target genes among top 100 ranked relationships, which indicates its potential important role in different diseases. This result can be verified to some extent by the reports in other literatures [71]-[73], e.g. knockdown of BANC remarkably promotes cell migration and proliferation of lung carcinoma through MAPK pathways [71], overexpression of BANC is related to clinical progression and can be used as prognostic biomarker in gastric cancer [72], downregulation of BANC makes contribution to the proliferation of colorectal cancer by means of downregulating p21 expression [73]. The similar conclusions of these researches highlight the importance of the mentioned lncRNAs and their relationships with genes, which may help make an improvement in the diagnosis and therapy of human cancers.

Although our approach achieves good performance and several findings are obtained in discovering relationships between lncRNAs and genes in human diseases, there are still some limitations in the generalization of this approach to other more diseases. For example, the model does not consider the directionality of the relationships between lncRNAs and genes, which is important in the analysis of disease progression. This issue will be addressed by modifying the model in the future study. Besides, the prediction cannot be performed on the diseases that have no prior information of known relationships between lncRNAs and genes due to the characteristics of tensor. In other words, a slice with no known entries in a tensor is not permitted in tensor completion based approach. Moreover, if the tensor has few known entries, e.g., the slice for each cancer has only one known entry [74], it would not be possible to faithfully recover the missing information in these extremely sparse data. However, the regulation of lncRNAs in diseases receives increasing attention and many related researches emerges [75]. In addition, benefit from the rapid development of biotechnology, more experimental data will be available. It is believed that a large number of comprehensive explorations are feasible in the future with the help of abundant known relationships between lncRNAs and genes. Certainly, any ternary relations that meet the requirement of data in this work can be studied by tensor completion based approach such as the relationships between miRNAs, target genes and cancers, which will also help us deepen the understand of progression of human diseases.

ACKNOWLEDGMENT
We would like to thank Wei-Li Guo from Tongji University for providing the information about tensor.

REFERENCES
a poor prognosis in non-small cell lung cancer and induces migration. Non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. TUG1 is overexpressed in urothelial carcinoma of bladder (Jung et al., 2011). Bulk sequence of TUG1 is downregulated in bladder cancer, whereas overexpression is found in bladder cancer (Monga et al., 2012).


Z. T. Chai, J. Kong, D. X. Zhu, Y. Zhang, L. Lu, J. M. Zhou et al., "MicroRNA-26a inhibits angiogenesis by down-regulating...


