Cooperative Driver Pathway Discovery by Hierarchical Clustering and Link Prediction

Sufang Li[1,2], Jun Wang[2,1,*], Maozu Guo[3], Xiangliang Zhang[4]
[1]College of Computer and Information Sciences, Southwest University, Chongqing, China
[2]Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan, China
[3]College of Elec. & Inf. Eng., Beijing University of Civil Engineering & Architecture, Beijing, China
[4]CEMSE, King Abdullah University of Science and Technology, Thuwal, SA
Email: sfli@email.swu.edu.cn; kingjun@sdu.edu.cn; guomaozu@bucea.edu.cn; xiangliang.zhang@kaust.edu.sa

***Abstract*—Identifying driver pathway is a critical step to uncover the natural laws of the occurrence and progression of disease. Many studies show that multiple pathways often function cooperatively in carcinogenesis. However, how to computationally identify cooperative driver pathways of cancers is not well studied yet. Existing cooperative driver pathway identification methods either suffer from single type of genetic information source or computation difficulty. In this paper, we proposed a method (CDPLP) based on hierarchical clustering and link prediction. CDPLP firstly devises a new similarity metric to quantify the exclusivity and co-expression of two gene modules, and thus to obtain gene sets with exclusivity by hierarchical clustering. Next, it uses link prediction on the pathway-pathway interaction network to replenish the interactions between pathways. After that, CDPLP combines the gene sets and updated pathway network to discover the pathway pairs with high functional interaction and occurrence as cooperative pathways. CDPLP can make full use of multiple genetic information sources such as the mutation data, gene-gene interaction data and pathway-pathway network, and facilitate the optimization solution. We evaluated the performance of CDPLP on TCGA breast cancer (BRCA) dataset and compared it with other popular methods. The results show that cooperative driver pathways identified by CDPLP are highly associated with the target cancer, and are involved with carcinogenesis and several key biological processes.**

***Index Terms*—Cooperative driver pathways; Cancer; Hierarchical clustering; Link prediction; Exclusivity and coverage**

I. Introduction

As a complex disease with high mortality rate [23], cancer has been attracting the attention of quite a few researchers. The ever-increasing incidence of cancer makes researchers eager to understand the pathogenesis of cancer. The previous study has indicted that mutations in a subset of genes that confer growth advantage result in cancers arising. To mine somatic mutation data, several projects (i.e., The Cancer Genome Atlas (TCGA) [19] and International Cancer Genome Consortium (ICGC) [8]) have been launched and they provide researchers with high-throughput data as expected. Previous researches have shown that only the functional driver mutations are important to cancer development while passenger mutations have less consequence for cancer [23]. To reduce the wet experimental cost and improve the capability of experimental validation in driver mutation discovery, several computational approaches have been proposed in the past years.

Most early efforts devoted to the detection of individual driver genes with significantly higher mutation rate [3]. For instance, MuSic [9] considers mutation types and sample-specific mutation rates. However, identifying and analyzing individual driver genes separately can not effectively explain the complex genetic mechanism of cancer. Recent studies have shown that driver pathways play a more important role in cancer development than individual driver genes, which consist of several driver genes and can control the progression of cancer from normal to malignant states.

High exclusivity (nearly all patients have no more than one mutation in candidate gene set.) is a key characteristic of a driver pathway and typically used in driver pathway discovery [1]. MEMo [7] builds a graph of all similar gene pairs, extracts all fully connected subgraphs and assesses each subgraph for mutual exclusivity to find driver module. Besides the high exclusivity, the *high coverage* (most patients have at least one mutation in the candidate gene set) is another key characteristic of a driver pathway. Dendrix [22] tries to find sets of genes with properties of high coverage and high exclusivity. To obtain these gene sets, Dendrix introduces a weight function to reward coverage while penalizing overlap to get high exclusivity. In this way, the objective can be converted into optimizing a maximum weight submatrix problem which is solved by Dendrix with Markov chain Monte Carlo (MCMC). However, Dendrix only gets an approximate solution and the stochastic search process may lead MCMC trapped into a local solution. To address this issue, MDPFinder [34] adopts a binary linear programming (BLP) model for efficient optimization to find the driver gene sets.

With the development of cancer researches, it has become a consensus that multiple driver pathways are cooperatively involved in the transformation process of a normal cell to tumor one during the cancer development [5]. The recent efforts are more devoted into the identification of multiple driver pathways. Multi-Dendrix [14] identifies multiple driver pathways by finding a collection of gene sets with a maximal sum of weights. It defines the problem of finding multiple driver pathways as a multiple maximum weight submatrix

*Corresponding author: kingjun@sdu.edu.cn (Jun Wang). This work is partially supported by NSFC (62072380, 62031003 and 61872300).

problem and solves it by the integer linear programming (ILP). However, multi-Dendrix may obtain only sub-optimal results and weaken the relations between pathways, since it neglects the functional interactions or co-occurrence of multiple pathways. Subsequently, CoMDP [31] was developed to discover co-occurring driver pathways with two properties: (1) each individual pathway has high coverage and high exclusivity; and (2) the mutations between the pair of pathways showed statistically significant co-occurrence. For the purpose, CoMDP defines a new weight function and solve it by a BLP model. CoMDP may miss some potential driver pathways since it identifies only super important genes with the highest coverage. Yang *et al.* [27] firstly integrated multiple prior knowledge data by matrix factorization and then applied a tri-random walk on a heterogeneous network composed with genes, miRNAs and pathways to identify cooperative driver pathways. However, this solution ignores mutation data, and suffers from too many parameters.

The cooperative driver pathway discovery methods mentioned above still have some limitations, such as the difficulty in optimizing process, neglecting pathway knowledge and locally optimal solutions. To overcome these limitations, we introduced a method named CDPLP to leverage prior knowledge and mutation profiles for cooperative driver pathways discovery. CDPLP firstly defines a new similarity metric based on the exclusivity and co-expression of genes for hierarchical clustering, which groups genes into gene sets with high coverage, mutual exclusivity and co-expression. Next, it leverage a heterogeneous network composed of three types of nodes (genes, miRNAs, pathways) and link prediction to replenish the potential associations between pathways. Finally, it maps the gene sets to pathway sets and identifies the pathway pairs with high functional interactions and co-occurrence as cooperative driver pathways. We apply CDPLP on TCGA somatic mutation profiles of breast cancer to identify cooperative driver pathways, each of which contains at least one reported driver gene and undertakes important conduction processes on the signaling network. The experimental results show that CDPLP is superior to the comparison methods in identifying driver genes which means that CDPLP can discover more potential cooperative pathways than existing competitive methods.

II. Methods

The whole procedure of CDPLP for discovering cooperative driver pathways is made up of three steps and illustrated in Figure 1. The following subsections will elaborate on these procedures.

A. The identification of driver gene modules

Driver genes tend to converge on a few biological driver pathways [26]. Thus, identifying driver gene modules is essential for the following driver pathway analysis. We firstly need to pick out the driver genes from many passenger ones. Recent researches indicate that a gene is more likely to represent a true cancer driver if it is functionally associated with other genes mutated in cancer [25], [29]. For these reasons, we adopt MUFFINN (MUtations For Functional Impact on Network Neighbors) [6] to fuse functional network and mutation data and to quantify the driver weight of each gene for picking out driver genes. MUFFINN prioritizes genes based on the mutation frequency of each gene as well as those of its neighbors in a functional network. For gene g, its mutation score $\omega(g)$ is calculated as follows:

$$\omega(g) = \frac{|\Gamma(g)|}{m} + \sum_{g' \in \mathcal{C}(g)} \frac{\omega(g')}{|\Gamma(g')|} \tag{1}$$

where $\Gamma(g)$ is the set of patients in which gene g mutates, $\mathcal{C}(g)$ is the set of neighborhood genes of gene g in two independently protein interaction network STRINGv11 and HumanNet [13], m is the number of all patients. In this way, each gene can obtain a mutation score to reflect how important it is. The top K genes will be retained to identify driver genes, while the other genes are excluded from follow-up analysis.

Clustering techniques are widely used for understanding gene functions, gene regulation, cellular processes, and subtypes of cells [12] [30]. Besides, hierarchical clustering (HAC) is free of the number of clusters and it is easy to define the distance. In view of this, we adopt the hierarchical agglomerative clustering on the selected gene candidate set to get diver gene modules. We introduce a new similarity metric for HAC to integrate the exclusivity and gene co-expression characteristics to quantify the proximity of two gene sets. For any two gene set $\mathcal{M}$ and $\mathcal{N}$, we define the similarity function as follows:

$$S(\mathcal{M},\mathcal{N}) = \alpha_1(1 - \frac{|T(\mathcal{M}) \cap T(\mathcal{N})|}{|T(\mathcal{M}) \cup T(\mathcal{N})|}) + \alpha_2 P(\mathcal{M},\mathcal{N}) \tag{2}$$

where $T(j)$ denotes the set of patients who have at least one mutated genes in gene set j, $j \in \{\mathcal{M},\mathcal{N}\}$. The two scalar parameters $\alpha_1, \alpha_2 \in (0,1)$ are employed to balance the contribution of exclusivity and gene co-expression. $1 - \frac{|T(\mathcal{M}) \cap T(\mathcal{N})|}{|T(\mathcal{M}) \cup T(\mathcal{N})|}$ is defined as the exclusivity of two gene sets. $P(\mathcal{M},\mathcal{N})$ quantifies the co-expression between two gene sets $\mathcal{M}$ and $\mathcal{N}$ and it is defined as follows:

$$P(\mathcal{M},\mathcal{N}) = \frac{1}{n_m * n_n} \sum_{g_m \in \mathcal{M}} \sum_{g_n \in \mathcal{N}} p(g_m, g_n) \tag{3}$$

where n_m (n_n) denotes the number of genes in gene set $\mathcal{M}$ ($\mathcal{N}$), and $p(g_m, g_n)$ is the Pearson correlation of gene g_m and gene g_n, which describes the similarity between expression patterns of the gene pair across all the samples.

Previous study indicates that a typical tumor contains two to eight of driver mutations [23], the clustering algorithm continuously merges two gene sets based on similarity scores calculated by Eq. (2) into a gene module until the number of genes in this module exceed eight. After the clustering, we can obtain mutually exclusive and co-expressed gene modules for the follow-up cooperative pathways identification.

B. The replenishment of pathway-pathway interactions

The information in pathway level also plays an important role in exploring the collaboration between pathways [25],

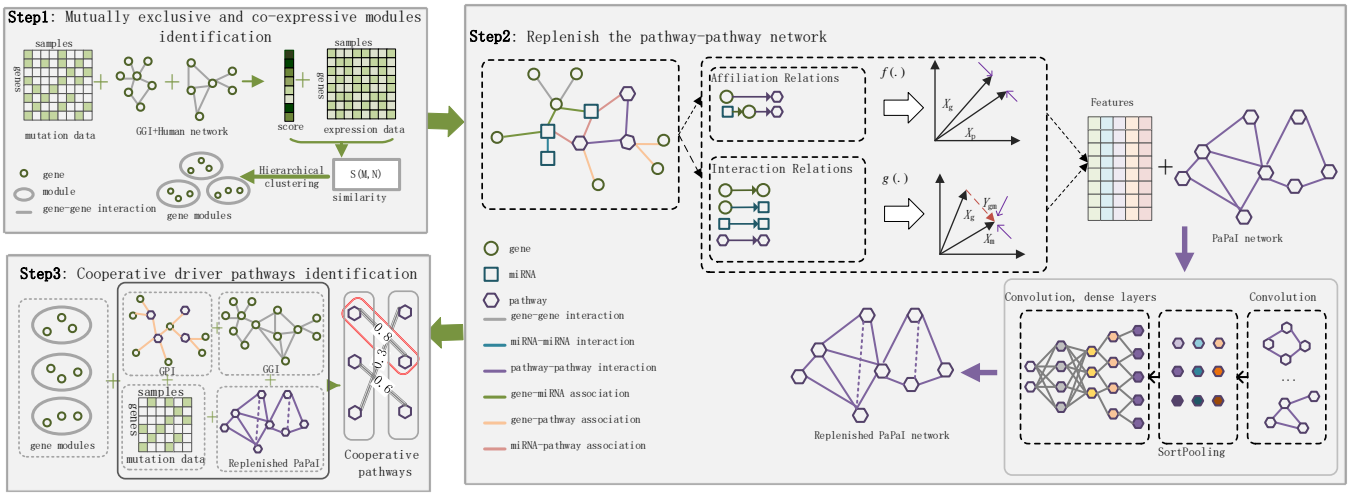

Fig. 1: The workflow of CDPLP. Step 1: CDPLP introduces a new similarity metric for hierarchical clustering to group genes into gene modules. Step 2: CDPLP applies link prediction on pathway-pathway interaction (PaPaI) network to replenish the interactions between pathways. Step 3: CDPLP defines a new cooperative quantification function based on the gene modules and updated PaPaI network to identify cooperative pathways.

[28], [29]. However, the known interactions between pathways are still incomplete. The interaction between signaling pathways is often responsible by the interactions between genes and miRNAs affiliated with the pathway, so it is critical to fully consider the relationship between genes, miRNAs and pathways to replenish the interactions between pathways. Link prediction is a popular method to predict whether two nodes in a network having a link or not and can be used to replenish the interactions between pathways in a biological network composed of pathway nodes. Given that, we adopt a representative link prediction method SEAL [32] to embed the features of nodes on the pathway-pathway interaction (PaPaI) network, predict the potential interactions between pathways and remedy the incompleteness of PaPaI network. The details are presented as follows.

1) Getting the pathway nodes features: We firstly construct a heterogeneous network $\mathbf{H}$ composed with nodes of genes, miRNAs, and pathways and the associations among them. Then, RHINE (Relation structure-aware Heterogeneous Information Network Embedding) [18] is employed to learn latent representations of nodes in our heterogeneous network. A degree based measure $D(r)$ is defined to explore the distinction of various relations in $\mathbf{H}$ and classifies the relations with a small value of $D(r)$ as *Interaction Relations* (IRs); otherwise, *Affiliation Relations* (ARs). $D(r)$ is defined as follows:

$$D(r) = \frac{\max[\bar{d}_{t_u}, \bar{d}_{t_v}]}{\min[\bar{d}_{t_u}, \bar{d}_{t_v}]} \tag{4}$$

where $\max[\bar{d}_{t_u}, \bar{d}_{t_v}]$ and $\min[\bar{d}_{t_u}, \bar{d}_{t_v}]$ are the maximum and minimum node degrees of types t_u and t_v (the node types include gene, miRNA and pathway), respectively.

According to the above Eq. (4), we divide relationships among nodes in $\mathbf{H}$ into ARs and IRs. Nodes connected by ARs share similar properties, therefore nodes could be directly close to each other in the vector space. IRs demonstrate strong interactions between nodes with compatible structural roles [18]. So, different proximity measures are used to quantify the distance between two nodes: euclidean distance for ARs and translation-based distance for IRs. Given an affiliation node-relation triple$< p, s, q > \in \mathcal{P}_{AR}$ with weight ω_{pg} and an interaction node-relation triple$< u, r, v > \in \mathcal{P}_{IR}$ with weight ω_{uv}, the distances of p, q and u, v are calculated as Eq. (5) and Eq. (6), respectively.

$$f(p,q) = \omega_{pq}||\mathbf{X}_p - \mathbf{X}_q||_2^2 \tag{5}$$

$$g(u,v) = \omega_{uv}||\mathbf{X}_u + \mathbf{Y}_r - \mathbf{X}_v|| \tag{6}$$

where $\mathbf{X}_p, \mathbf{X}_q, \mathbf{X}_u, \mathbf{X}_v \in \mathbb{R}^c$ are the embedding vectors of p, q, u and v, respectively. $\mathbf{Y}_r$ is the embedding of the relation r. To ensure that nodes connected by an AR should be close to each other, a margin-based loss for ARs is defined in Eq. (7). For the same reason, a margin-based loss for IRs is defined as:

$$L_{EuAR} = \sum_{s \in \mathcal{R}_{AR}} \sum_{<p,s,q> \in \mathcal{P}_{AR}} \sum_{<p',s,q'> \in \mathcal{P'}_{AR}} \max[0, \gamma + f(p,q) - f(p',q')] \tag{7}$$

$$L_{TrIR} = \sum_{\gamma \in \mathcal{R}_{IR}} \sum_{<u,\gamma,v> \in \mathcal{P}_{IR}} \sum_{<u',\gamma,v'> \in \mathcal{P'}_{IR}} \max[0, \gamma + g(u,v) - g(u',v')] \tag{8}$$

where $\gamma > 0$ is a margin hyperparameter. $\mathcal{P}_{AR}$ and $\mathcal{P}_{IR}$ are the set of positive affiliation and interaction node-relation triples, while $\mathcal{P'}_{AR}$ and $\mathcal{P'}_{IR}$ are the set of negative ones respectively.

Consider that there are both ARs and IRs in $\mathbf{H}$, a unified model for $\mathbf{H}$ embedding is defined as follows:

$$L = L_{EuAR} + L_{TrIR} \tag{9}$$

By minimizing the above equation, the embedding of $\mathbf{H}$ is obtained. After that, we pick out the embedding of pathway nodes from the whole embedding and splice them into matrices. In this way, we get the PaPaI network information matrix $\mathbf{X} \in \mathbb{R}^{n \times c}$ with n pathways and c feature channels.

2) Getting the enclosing subgraphs features: Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the PaPaI network. The node feature matrix $\mathbf{X} \in \mathbb{R}^{n \times c}$ of the PaPaI network was obtained in the previous step, the graph convolution layer takes the following form:

$$\mathbf{Z} = f(\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{A}}\mathbf{X}\mathbf{W}) \tag{10}$$

where $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix of the graph with added self-loops, $\widetilde{\mathbf{D}}$ is its diagonal degree matrix. $\mathbf{W} \in \mathbb{R}^{c \times c'}$ is a matrix of trainable graph convolution parameters, f is a nonlinear activation function, and $\mathbf{Z} \in \mathbb{R}^{n \times c'}$ is the output activation matrix. The graph convolution maps the c features to c' features by a linear feature transformation on node information matrix $\mathbf{X}$. Then it propagates node information to neighborhood vertices as well as itself. After that, it normalizes the features by multiplying $\widetilde{\mathbf{D}}^{-1}$. Finally, it applies a pointwise nonlinear activation function f and outputs the subgraph features.

After the above procedure, we get all subgraph features of PaPaI network. We then input the subgraph features and the positive interactions between pathways to DGCNN [33] to train a link prediction model. Next, we input all the negative interactions as unknown interactions and apply DGCNN to identify additional interactions between pathways. We get the probability of the positive interactions. We deem them as connected when the probability between the two pathways exceeds the threshold (0.5). After that, we combine the identified interactions and original PaPaIs to get a update PaPaI network.

C. Identifying cooperative pathways

Based on the gene modules and replenished pathway network, we define a new quantitative function to discover cooperative driver pathways. In the previous efforts [25], [28], [29], high co-occurrence and large functional cooperation between genes of pathways are regarded as important criteria for the identification of cooperative pathways. However, the existing interaction between pathways is also an important factor for cooperative pathway discovery. Given that, we introduce the updated interactions between pathways and define a quantitative function to evaluate the cooperation between two pathways (p_1 and p_2) as follows:

$$\begin{aligned} Co(p_1, p_2) = \gamma_1 \sum_{g_1 \in \Theta(p_1)} \sum_{g_2 \in \Theta(p_2)} I(g_1, g_2) \\ + \gamma_2 O(p_1, p_2) + \gamma_3 Pac(p_1, p_2) \end{aligned} \tag{11}$$

where $I(g_1, g_2)$ is the interaction score between gene g_1 and g_2 in the GGI network, $O(p_1, p_2)$ denotes the total number of samples that genes in pathway p_1 and p_2 both mutated, and $Pac(p_1, p_2)$ is the interaction score of p_1 and p_2 in PaPaI network. $\Theta(p_i)$ includes the genes related to p_i, $i \in \{1, 2\}$. The three scalar parameters γ_1, γ_2 and $\gamma_3 \in (0, 1)$ are employed to balance the contribution of gene interactions, gene co-occurrence and pathway interactions, respectively. We calculate the cooperation score of the each pathway pairs based on Eq. (7), and take the top ten pathway pairs with the highest cooperation score as cooperative pathways.

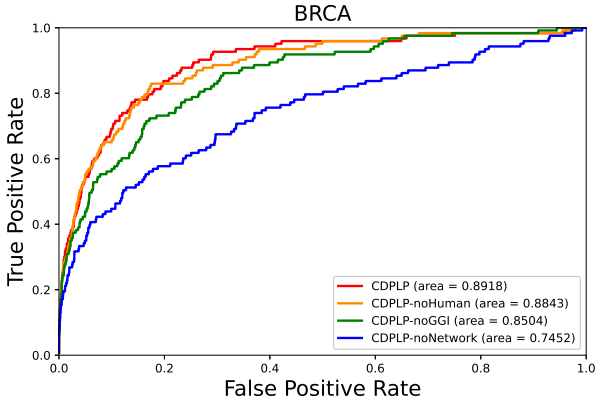

Fig. 2: Receiver operating characteristics (ROC) curves of identifying known driver genes on BRCA . The corresponding AUROC values are also included in the legend. CDPLP-noGGI: CDPLP without the GGI network. CDPLP-noHuman: CDPLP without the HumanNet

III. RESULTS

A. Data sources and preprocessing

To investigate the effectiveness of CDPLP, we carried out experiments on publicly available cancer data of breast cancer (BRCA) from TCGA [19]. It consists of the raw mutation data (with 791 samples and 40543 genes) and gene expression data (with 1218 samples and 20530 genes). We selected the samples and genes both in the mutation and expression data to generate a dataset (with 789 samples and 18616 genes) for BRCA. We obtained the disease related gene-miRNA networks from HMDD [15], gene-gene interaction network from [11] and miRNA-miRNA interaction network from [24]. We downloaded PaPaI network data and gene-pathway association from PID database [21]. In summary, we construct a heterogeneous network with 19342 genes, 559 miRNAs and 212 pathways.

B. Results on real data

We apply CDPLP on the BRCA dataset to evaluate the effectiveness. We determine the hype-parameters by grid search in 0 to 1. The parameters α_1, α_2, γ_1, γ_2 and γ_3 are finally set as 0.7, 0.8, 0.7, 0.8 and 0.3. According to the analysis of K, we set $K = 6000$. We have the following conclusions.

1) CDPLP can effectively identify gene modules: CDPLP selects the genes with top K mutation scores as the driver gene candidates according Eq. (1). We analyze the change of K (selected candidate genes) with respect to the truth driver genes. According to the analysis results, we pick the top 6000 genes as driver gene candidates and exclude the other genes from follow-up analysis. To investigate the contribution of GGI

and HumanNet, we also compare CDPLP with CDPLP without the GGI or HumanNet. When we disregard GGI network, the AUROC value are reduced by 4.1 % on BRCA. When both networks are removed, the AUROC value drops sharply by 14.7% on BRCA. The result can prove the feasibility of merging the GGI and HumanNet in driver genes identification.

We cluster candidate driver genes into gene modules according to Eq. (2). To verify that the identified driver gene modules are with high similarity, we randomly generate 1000 gene sets with 8 genes and calculate the similarity of them according to Eq. (2). Then, we calculate the average similarity of the identified gene sets and compare it with the random similarity. The results indicate that the similarity of gene modules identified by CDPLP is over 98.6% than that of random gene set for BRCA dataset. Given that, we can conclude that CDPLP can identify high mutual exclusivity and co-expression gene modules.

In addition, we collect the genes related with breast cancer from DisGeNet [20], and estimate the relation between gene modules and target cancer. About 91% gene modules include at least three genes that are associated with breast cancer. To further validate the functions of gene modules, we do GO and KEGG enrichment analysis on STRING for all modules. As shown in Table I, the analysis of TP53 gene module show associations with breast cancer and with several carcinogenesis related activities, such as cell cycle, cell migration and so on. Enriched pathways, including PI3K and ErbB, have been reported to be related with breast cancer.

Based on the above analysis on the identified gene modules, we can conclude that CDPLP can effectively identify driver gene modules, which have close associations with the target cancer and important biological activities involved with cancer development.

2) CDPLP can effectively identify cooperative driver pathways: For BRCA dataset, we find Class I PI3K signaling events, ATM pathway, Atypical NF-kappaB pathway, ErbB1 downstream signaling and their cooperative pathways. It is recognized that the occurrence, development, metastasis and drug resistance of breast cancer are closely related to intracellular signaling pathways, including insulin-like growth factor receptor (IGFR) signaling pathway and epidermal growth factor receptor (EGFR) signaling pathway, all of which are particularly important [10]. For Atypical NF-kappaB pathway, its inhibitors preferentially inhibit breast cancer stem-like cells [35]. As for Class I PI3K signaling events, it is altered in a high proportion of breast cancers and may contribute to therapeutic resistance [17].

Based on the above analysis, we can state that CDPLP can effectively identify the target cancer related cooperative pathways, and the cooperation between them can be further investigated by wet experiments.

3) Comparison with other methods: To comparatively study the effectiveness of CDPLP, we compare it against four related and competitive methods, including an individual driver pathways identification method: Dendrix [22], and three multiple driver pathways identification methods: CoMDP [31], Multi-Dendrix [14], CDPathway [25]. We set the parameter ranges of compared methods as described in the original papers, and select pathways that contain the most known driver genes with statistic significance.

TABLE I: Functional analysis of identified module on BRCA

Category	Terms	Genes
GO	cell surface receptor signaling pathway	TP53, UBB, AKT1 HRAS, RHOA, MAPK3, CCND1;
GO	cell cycle	TP53, HRAS, RHOAMAPK3 CCND1, RHOC;
GO	regulation of organelle organization	TP53, RHOC, AKT1 HRAS, RHOA, MAPK3;
GO	positive regulation of cell migration	RHOC, AKT1, HRAS RHOA, MAPK3;
KEGG	Breast cancer	TP53, AKT1, HRAS MAPK3, CCND1;
KEGG	ErbB signaling pathway	AKT1, HRAS, MAPK3;
KEGG	PI3K-Akt signaling pathway	TP53, HRAS, CCND1 AKT1, MAPK3;
KEGG	Ras signaling pathway	AKT1, HRAS, MAPK3, RHOA;

TABLE II: Driver pathways of BRCA identified by CDPLP and other compared methods.

Method	Driver Pathways of BRCA
CDPLP	Atypical NF-kappaB pathway, p53 pathway Direct p53 effectors, EGF receptor (ErbB1) signaling pathway ErbB1 downstream signaling, Arf6 downstream pathway Class I PI3K signaling events, FoxO family signaling C-MYB transcription factor network, Aurora A signaling ATM pathway, Class I PI3K signaling events mediated by Akt
CDPathway	p53 pathway, Class I PI3K signaling events mediated by Akt, Internalization of ErbB1 Alternative NF-kappB pathway, Trk recptor signaling mediated by PI3K and PLC-gamma, Degradation of beta catenin Direct p53 effectors, FAS (CD95) signaling pathway Fanconi anemia pathway, Aurora A signaling C-MYC pathway, ErbB1 downstream signaling
Dendrix	RAC1 signaling pathway, Signaling events mediated by CDC42 signaling events Hepatocyte Growth Factor Receptor (c-Met)
Multi-Dendrix	BARC1 signaling pathway, Class I PI3K signaling events
CoMDP	ErbB1 downstream signaling, p53 pathway p75(NTR)-mediated signaling Signaling events mediated by Hepatocyte Growth Factor (c-Met)

We compare CDPLP with other methods at the pathway level. As shown in Table II, pathways identified by CDPLP cover more breast cancer related pathways. For example, Class I PI3K signaling events [17], ErbB1 downstream signaling [2] and Atypical NF-kappaB pathway [35] are reported to related with breast cancer. In addition, CDPLP has overlapped pathways with 3/4 compared methods on BRCA dataset. Specially, the number of driver pathways identified by CDPLP and CDPathway are similar and the two methods have a large overlap. However, CDPLP discovers more reported driver pathways than CDPathway. For example, [16] reported that C-MYB transcription factor network enhances breast cancer invasion and metastasis and the overexpression of the erbB-1 (EGFR, epidermal growth factor receptor) proteins contributes to the aggressive behavior of malignant tumors originating from the enedometrium [4]. Compared with CDPathway, CDPLP focuses only on the main factors (the influence of neighbor genes on candidate genes), so CDPLP reduces the calculation consumption at the cost of a slightly reduced accuracy. Therefore, CDPLP can not only identify more driver pathways than individual methods but also uncover more

potential cooperations between pathways.

IV. Conclusion

The cooperative pathways discovery can help us to understand the genetic mechanisms of complex diseases, and to provide more effective genetic therapy to patients. In this paper, we proposed a novel method named CDPLP, which combines biological data with link prediction algorithms to identify cooperative pathway. The main contribution of CDPLP is that it firstly applies network representation learning method to the task of cooperative pathways identification and defines a new metric to quantify the cooperation of pathways. Experimental results on public breast cancer dataset show that CDPLP can identify more potential biological cooperation among pathways. The future pursue for CDPLP is to reduce the number of input parameters.

References

[1] O. Babur, M. Gonen, B. A. Aksoy, N. Schultz, G. Ciriello, C. Sander, and E. Demir, "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations," *Genome Biology*, vol. 16, no. 1, pp. 45–45, 2015.

[2] S. P. Bagaria, P. S. Ray, M.-S. Sim, X. Ye, J. M. Shamonki, X. Cui, and A. E. Giuliano, "Personalizing breast cancer staging by the inclusion of er, pr, and her2," *JAMA Surgery*, vol. 149, no. 2, pp. 125–129, 2014.

[3] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander *et al.*, "Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 20 007–20 012, 2007.

[4] M. Brys, A. Semczuk, T. Rechberger, and W. M. Krajewska, "Expression of erbb-1 and erbb-2 genes in normal and pathological human endometrium," *Oncology Reports*, vol. 18, no. 1, pp. 261–265, 2007.

[5] J. Chen and S. Zhang, "Integrative cancer genomics: models, algorithms and analysis," *Frontiers of Computer Science*, vol. 11, no. 3, pp. 392–406, 2017.

[6] A. Cho, J. E. Shim, E. Kim, F. Supek, B. Lehner, and I. Lee, "Muffinn: cancer gene discovery via network analysis of somatic mutation data," *Genome Biology*, vol. 17, no. 1, pp. 129–129, 2016.

[7] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome Research*, vol. 22, no. 2, pp. 398–406, 2012.

[8] I. C. G. Consortium, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, p. 993, 2010.

[9] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. J. Dooling, E. R. Mardis *et al.*, "Music: Identifying mutational significance in cancer genomes," *Genome Research*, vol. 22, no. 8, pp. 1589–1598, 2012.

[10] M. Elbaz, M. W. Nasser, J. Ravi, N. A. Wani, D. K. Ahirwar, H. Zhao, S. Oghumu, A. R. Satoskar, K. Shilo, W. E. Carson III *et al.*, "Modulation of the tumor microenvironment and inhibition of egf/egfr pathway: Novel anti-tumor mechanisms of cannabidiol in breast cancer," *Molecular Oncology*, vol. 9, no. 4, pp. 906–919, 2015.

[11] Y. Hou, B. Gao, G. Li, and Z. Su, "Maxmif: a new method for identifying cancer driver genes through effective data integration," *Advanced Science*, vol. 5, no. 9, p. 1800640, 2018.

[12] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[13] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome research*, vol. 21, no. 7, pp. 1109–1121, 2011.

[14] M. D. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, "Simultaneous identification of multiple driver pathways in cancer," *PLoS Computational Biology*, vol. 9, no. 5, p. e1003054, 2013.

[15] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "Hmdd v2.0: a database for experimentally supported human microrna and disease associations," *Nucleic Acids Research*, vol. 42, pp. 1070–1074, 2014.

[16] Y. Li, K. Jin, G. W. van Pelt, H. van Dam, X. Yu, W. E. Mesker, P. ten Dijke, F. Zhou, and L. Zhang, "c-myb enhances breast cancer invasion and metastasis through the wnt/β-catenin/axin2 pathway," *Cancer Research*, vol. 76, no. 11, pp. 3364–3375, 2016.

[17] E. López-Knowles, S. A. O'Toole, C. M. McNeil, E. K. Millar, M. R. Qiu, P. Crea, R. J. Daly, E. A. Musgrove, and R. L. Sutherland, "Pi3k pathway activation in breast cancer is associated with the basal-like phenotype and cancer-specific mortality," *International Journal of Cancer*, vol. 126, no. 5, pp. 1121–1131, 2010.

[18] Y. Lu, C. Shi, L. Hu, and Z. Liu, "Relation structure-aware heterogeneous information network embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4456–4463.

[19] C. G. A. R. Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, p. 609, 2011.

[20] J. Pinero, J. M. Ramirezanguita, J. Sauchpitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update." *Nucleic Acids Research*, vol. 48, no. D1, pp. D845–D855, 2020.

[21] C. F. Schaefer, K. Anthony, S. Krupa, J. R. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "Pid: the pathway interaction database," *Nucleic Acids Research*, vol. 37, pp. 674–679, 2009.

[22] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," *Genome Research*, vol. 22, no. 2, pp. 375–385, 2012.

[23] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes." *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[24] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microrna functional similarity and functional network based on microrna-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[25] J. Wang, Z. Yang, C. Domeniconi, X. Zhang, and G. Yu, "Cooperative driver pathway discovery via fusion of multi-relational data of genes, mirnas and pathways." *Briefings in Bioinformatics*, vol. 00, no. 00, pp. 1–17, 2020.

[26] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak *et al.*, "The genomic landscapes of human breast and colorectal cancers." *Science*, vol. 318, no. 5853, pp. 1108–1113, 2007.

[27] Z. Yang, G. Yu, M. Guo, and J. Wang, "Codp: cooperative driver pathways discovery with matrix factorization and tri-random walk," *IEEE Access*, vol. 7, pp. 77 738–77 749, 2019.

[28] Z. Yang, G. Yu, M. Guo, J. Yu, X. Zhang, and J. Wang, "Cdpath: Cooperative driver pathways discovery using integer linear programming and markov clustering," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 99, no. 99, pp. 1–11, 2019.

[29] Z. Yang, G. Yu, J. Yu, M. Guo, and J. Wang, "Copath: discovering cooperative driver pathways using greedy mutual exclusivity and bi-clustering," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2019, pp. 165–170.

[30] X. Yu, G. Yu, J. Wang, and C. Domeniconi, "Co-clustering ensembles based on multiple relevance measures," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. 99, pp. 1–12, 2020.

[31] J. Zhang, L. Wu, X. Zhang, and S. Zhang, "Discovery of co-occurring driver pathways in cancer," *BMC Bioinformatics*, vol. 15, no. 1, p. 271, 2014.

[32] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 5165–5175.

[33] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 4438–4445.

[34] J. Zhao, S. Zhang, L. Wu, and X. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," *Bioinformatics*, vol. 28, no. 22, pp. 2940–2947, 2012.

[35] J. Zhou, H. Zhang, P. Gu, J. Bai, J. B. Margolick, and Y. Zhang, "Nf-κb pathway inhibitors preferentially inhibit breast cancer stem-like cells," *Breast Cancer Research and Treatment*, vol. 111, no. 3, pp. 419–427, 2008.