Systems biology

Cooperative driver pathway discovery via fusion of multi-relational data of genes, miRNAs, and pathways

Jun Wang¹, Ziyang Yang¹, Carlotta Domeniconi², Xiangliang Zhang³, and Guoxian Yu^{1, *}

¹College of Computer and Information Sciences, Southwest University, Chongqing 400715, China.

²Department of Computer Science, George Mason University, Fairfax 22030, VA, USA.

³CEMSE, King Abdullah University of Science and Technology, Thuwal, SA.

Corresponding Author: gxyu@swu.edu.cn (Guoxian Yu).

Abstract

Discovering driver pathways is an essential step to uncover the molecular mechanism underlying cancer, and to explore precise treatments for cancer patients. However, due to the difficulties of mapping genes to pathways, and the limited knowledge about pathway interactions, most previous work focus on identifying individual pathways. In practice, two (or even more) pathways interplay and often cooperatively trigger cancer. In this study, we proposed a new approach called CDPPathway to discover cooperative driver pathways. First, CDPPathway introduces a driver impact quantification function to quantify the driver weight of each gene. CDPPathway assumes genes with larger weights contribute more to the occurrence of the target disease and identifies them as candidate driver genes. Next, it constructs a heterogeneous network composed of genes, miRNAs, and pathways nodes based on the known intra(inter)-relations between them, and assigns the quantified driver weights to gene-pathway and gene-miRNA relational edges. To transfer driver impacts of genes to pathway interaction pairs, CDPPathway collaboratively factorizes the weighted adjacency matrices of the heterogeneous network to explore the latent relations between genes, miRNAs, and pathways. After this, it reconstructs the pathway interaction network, and identifies the pathway pairs with maximal interactive and driver weights as cooperative driver pathways. Experimental results on the breast, uterine corpus endometrial carcinoma and ovarian cancer data from TCGA show that CDPPathway can effectively identify candidate driver genes ($AUROC \geq 0.9$) and reconstruct the pathway interaction network ($AUROC > 0.9$); and it uncovers much more known (potential) driver genes than other competitive methods. In addition, CDPPathway identifies 150% more driver pathways and 60% more potential cooperative driver pathways than the competing methods. The code of CDPPathway is available at <http://mla.swu.edu.cn/codes.php?name=CDPPathway>.

Key words: cooperative driver pathway, cancer gene, microRNA, data fusion, biological network

***Jun Wang** is an Associate Professor of College of Computer and Information Sciences, Southwest University. Her research interests include cancer data mining, pathway analysis, GWAS and machine learning.

Ziyang Yang is a Master student of College of Computer and Information Sciences, Southwest University. Her research interests include cancer pathways analysis and biological data fusion.

Carlotta Domeniconi is an Associate Professor of Department of Computer Science, George Mason University. Her research interests

include biological data mining and machine learning.

Xiangliang Zhang is an Associate Professor of Computer Science, King Abdullah University of Science and Technology, SA. Her research interests include data mining and machine learning.

Guoxian Yu is a Professor of College of Computer and Information Sciences, Southwest University. His research interests include gene function prediction, ontology analysis and data mining.

1 Introduction

Cancer is one of the most complex diseases that threaten human health. Systemic cancer genomics projects, such as the Cancer Genome Atlas (TCGA) [1] and the International Cancer Genome Consortium (ICGC) [2], have taken initial steps toward developing a blueprint of human cancer genomes by identifying, characterizing, and cataloguing alterations in thousands of tumor samples. A major challenge of interpreting the data in these projects is to distinguish the mutations that play a role in the initiation and progression of cancer, from the much larger number of passenger alterations that have little influence in cancer cell development.

Most of the early studies focus on detecting individual driver genes based on periodic mutation rates of genes in a large population of cancer patients [3, 4]. The useful implementation of this concept, however, faces a number of well known challenges. First, the vast majority of driver genes occur rarely, making them difficult to be detected statistically [5]. Second, different types of mutations exist and the amount of mutated genes is very large; as a consequence, the sets of genes involved in different studies typically have a small overlap. Therefore, it is difficult to establish a consistent causal mechanism for a given cancer [6]. Third, transformed cells are typically mutated in multiple members of a set of functionally related genes. Consequently, mutations that drive transformation, are best sought and understood in a pathway context, especially when the mutation rates are relatively low [7]. For these reasons, efforts have been moving toward identifying driver pathways [8, 9, 10].

One way to proceed is to identify known pathways that are enriched in genes carrying somatic mutations [1, 4]. More recent methods exploit genomic characteristics of mutations (i.e., mutual exclusivity that genes in the same pathway rarely mutate in the same sample), to identify oncogenic modules [11, 12]. Vandin *et al.* [13] proposed Dendrix (*de novo* driver exclusivity), which maximizes a novel weight function that considers the high coverage and mutual exclusivity of the driver gene set, to discover mutated driver pathways. To better address the maximum weight function, Zhao *et al.* [14] proposed an exact binary linear programming (BLP) model to maximize the weight function. Zhang *et al.* [15] proposed two optimization models to separately discover common driver gene sets among multiple cancer types (ComMDP), and specific driver gene sets of a particular cancer (SpeMDP). However, these *de novo* methods focus on single pathways or modules, and they cannot shed light on how various cellular and physiological processes cooperatively alter during the initiation and progression of cancer. Besides mutation data, gene level knowledge including gene size bias [16, 17], gene-interaction networks, and expression levels [5, 7, 18] have been incorporated to uncover the significant mutations of a pathway. Copy number alterations [19] and other biological knowledge of mutational processes, such as transcript isoforms, variation in mutation type and redundancy of genetic code, were also integrated for analysis [20]. Approaches that infer patient specific pathways have also been developed [21, 22]. However, these methods ignore pathway interactions and cannot directly identify driver pathways. Furthermore, all the aforementioned methods may be biased towards super important genes, while ignoring other important driver genes with a low mutation rate.

It has been recognized that pathways often function cooperatively in carcinogenesis [23, 24]. For example, TGF- β (transforming growth factor- β) transactivates EGFR (epidermal growth factor receptor) through canonical Smad3 and ERK/Sp1 signaling pathways, and then facilitates breast cancer migration and invasion [25]. Unfortunately, computational solutions for cooperative driver pathway discovery are still much limited, due to the difficulty of mapping genes to pathways and the incomplete pathway knowledge. Furthermore, it is difficult to collect the information of molecules within pathways and the associations among pathways. Yeang *et al.* [12] reported that there are significant combinatorial patterns

of mutations occurring in the same patients (i.e., *co-occurrence*), whose driver genes usually function in different pathways, and genes in the same pathway rarely mutate in the same sample (i.e., *mutual exclusivity*). Leiserson *et al.* [26] proposed Multi-Dendrix, which uses a sum of multiple quantities for multiple pathways to identify driver pathways. Multi-Dendrix may obtain only sub-optimal results and weaken the relations between pathways, since it does not consider the functional interactions or co-occurrence of multiple pathways. Zhang *et al.* [27] proposed CoMDP, which extends Dendrix by maximizing the mutual exclusivity within a pathway and the co-occurrence of mutations between different pathways. But CoMDP can only obtain one pair of cooperative driver gene sets for a target disease. So CoMDP can be easily trapped, only identify super important genes with the highest coverage and miss other potential driver pathways. Dao *et al.* [28] proposed BeCo-WithMEFun to discover cooperative driver pathways. BeCo-WithMEFun pursues the mutual exclusivity and functional interactions within pathways, and the co-occurrence between pathways using Integer Linear Programming (ILP). However, when there are insufficient co-occurrences or functional interactions among genes, BeCo-WithMEFun cannot find a solution, or at best can only find few genes. This is because gene modules with joint mutual exclusivity, co-occurrence, and functional interaction are rare. Yang *et al.* [29] first integrated multiple prior knowledge data by matrix factorization and then applied a tri-random walk to identify cooperative driver pathways. This solution ignores mutation data, and it suffers from too many parameters and high time cost.

Existing cooperative driver pathway discovery methods suffer from at least one of the following issues: neglecting pathway knowledge; unable to directly identify cooperative driver pathways [26, 27, 28]; limited information given by (noisy) single type of data [30]; difficulty in optimizing the problem [28]; or tedious parameter selection [29]. To address these issues, we develop a novel approach called CDPPathway to leverage prior knowledge (i.e., gene-gene interactions and signal networks) and mutation profiles for cooperative driver pathways discovery. CDPPathway identifies candidate driver genes based on a driver impact quantification function that integrates somatic mutation, signaling network, and gene-gene interaction (GGI) data. The quantification function measures the importance score of each gene in the mutation profile and prior functional networks, and then takes the score as driver weight of the gene. In this way, CDPPathway can uncover driver genes with maximal weights that play important roles in the carcinogenesis and signaling network. Next, CDPPathway imports the quantified driver weights, along with the multi-relational data of genes, miRNAs and pathways, to construct a heterogeneous network, in which the edges of gene-miRNA and gene-pathway are specified by the quantified driver weights. To map driver genes to driver pathways and to uncover the cooperations between these driver pathways, CDPPathway collaboratively factorizes the weight adjacency matrices of the heterogeneous network. As such, CDPPathway can transfer the influence of driver genes on disease to pathways, to reconstruct the pathway interaction network. After that, CDPPathway identifies the interactive pathways with the maximal driver weights on disease as cooperative driver pathways.

We apply CDPPathway on two real TCGA datasets: somatic mutation and copy number variation profiles of breast cancer (BRCA) and of uterine corpus endometrial carcinoma (UCEC). After statistical and biological enrichment analysis (GO biological process terms and KEGG pathways), the genes identified by the driver impact quantification function show strong associations with disease (AUROC ≥ 0.9) and play crucial roles in important biological processes. In addition, these identified genes are tightly connected in signaling pathways, which suggests that cooperations do exist between pathways to cooperatively trigger cancer. We also evaluate the effectiveness of the collaborative matrix factorization via five fold cross validation, and obtain an AUROC > 0.9 , which supports the

Fig. 1. The workflow of CDPATHWAY. Step 1: CDPATHWAY introduces a driver impact quantification function to measure the importance score of each gene to disease based on the mutation profile and on prior functional networks integrated by GGI and signaling networks, and takes the score as driver weight of each gene. Step 2: CDPATHWAY imports the driver weights to construct a heterogeneous network composed of inter-relational and intra-relational networks of genes, miRNAs, and pathways, and then collaboratively factorizes the weight adjacency matrices of the heterogeneous network to reconstruct the pathway-to-pathway interaction network. CDPATHWAY identifies the pathway pairs with the maximal interactive and driver weight as cooperative driver pathways.

credibility of the identified pathway cooperations. Finally, we evaluate the identified cooperative driver pathways. We find that 100% of them contain driver genes and are related to important biological activities. We compare CDPATHWAY with related methods, including individual driver pathways identification methods: Dendrix [13] and BLP [14]; and multiple driver pathways identification methods: CoMDP [27], Multi-Dendrix [26], BeCo-WithMEFun [28] and CoDP [29]. At gene level, CDPATHWAY can identify a larger number of known driver genes with higher statistical significance. At pathway level, CDPATHWAY performs better by 150% in identifying more driver pathways and by 60% in uncovering more potential co-operations between driver pathways than competitive methods. **We also apply CDPATHWAY on an ovarian cancer dataset (8911 genes $\times$ 316 genes, detailed results are shown in Section 3 of the Supplementary file). Results show that CDPATHWAY can identify more known driver genes by over 100%, and uncover more significant cooperations between pathways by 80%, which further show the effectiveness of CDPATHWAY on different datasets.**

2 Methods

To identify cooperative driver pathways, CDPATHWAY follows the two-step procedure illustrated in Figure 1.

2.1 Step 1: Identify candidate driver genes via a driver impact quantification function

Mutations in a cancer genome tend to converge on a few biological pathways [31] and there are relatively complete mutation data on diseases. Therefore, we first identify candidate driver genes in this step and then map these candidates to pathways. In this way, the effects of the candidate driver genes on cancer can be transferred to pathways. In addition, pathway-based or network-based approaches [21, 22] for discovering cancer genes have shown that functional networks are helpful to identify cancer driver genes. Therefore, we import the gene-gene interaction (GGI) network and signaling network to identify driver genes. Several methods have been introduced to integrate mutation and prior knowledge data to identify driver

genes [32, 33]. Inspired by MaxMIF [33], we introduce a new driver impact quantification function to quantify the driver weight of each gene and to identify candidate driver genes using the weights.

Previous studies indicate that a driver gene set has the key property of covering a large number of samples (high *coverage*) [13, 28]. As such, we calculate the mutation score for each gene to measure the contribution of its mutations to cancer. The mutation data is stored in a binary mutation matrix $\mathbf{X}$, in which the rows represent the genes, and the columns represent the cancer samples (patients). For a protein-coding gene i , $\mathbf{X}(i, k) = 1$ if it has at least one nonsilent somatic mutation in sample k . For a non-protein-coding gene i , $\mathbf{X}(i, k) = 1$ if it has at least one mutation; and $\mathbf{X}(i, k) = 0$ otherwise. The mutation score $\mathbf{M}(i)$ for each gene i is computed as:

$$\mathbf{M}(i) = \begin{cases} \sum_{k \in \mathcal{K}_i} \frac{1}{N_k}, & \mathcal{K}_i \neq \emptyset \\ 1/N_{max}, & \mathcal{K}_i = \emptyset \end{cases} \quad (1)$$

where $\mathcal{K}_i$ is the set of samples in which gene i mutates, N_k is the total number of mutated genes in sample k , and N_{max} is the maximal number of mutated genes in all the samples. If gene i mutates in the samples, the mutation score of gene i is ensured to be equally affected by the samples in which gene i mutates, regardless of the total number of mutated genes in the samples. In this way, Eq. (1) can balance the contribution of all the samples with different numbers of mutated genes. If gene i does not mutate in any sample, $\mathbf{M}(i)$ is defined as a background mutation score (BMS), which is no larger than the mutation score of any other mutated genes. In this way, BMS can help identifying possible driver genes with an extremely low mutation rate, but that play important roles in functional networks. Therefore, driver genes with a small number of mutations can still be discovered.

The gravity model derived from Newton’s law of gravitation has been used in several fields to calculate the gravitation of two bodies, e.g., population migration [34]. **To further import the biological functions of genes, we assume that the genetic interaction between genes i and j follows a gravity model. The gene gravity model assumes that if two genes have high mutation density and strong gene co-expression in a given cancer type,**

they should have a higher score and related to a higher risk of inducing mutations to other genes [35, 36, 37, 38]. Then we measure the impact of interactions between two mutated genes based on biological functions, and calculate the functional mutation score of two genes i and j as follows:

$$\mathbf{F}(i, j) = \mathbf{M}(i)\mathbf{M}(j)/r_{ij}^2 \quad (2)$$

where $\mathbf{M}(i)$ is the mutation score of gene i computed by Eq. (1), and $r_{ij} = 1/\mathbf{W}(i, j)$ is the ‘interaction distance’ between genes i and j , $\mathbf{W}(i, j)$ is the average interaction weight between genes i and j in the GGI and signaling networks. The interaction weights in the GGI and signaling networks indicate the confidence or interaction strength of two genes. If the i - j gene pair is only available in the GGI or signaling network, the interaction weight between them in the unavailable network is 0. Thus, $\mathbf{F}$ can integrate mutation information and functional relationships between two genes. Two genes with high mutation rates and (or) interacting tightly in the GGI and signaling networks have a high functional mutation value.

By applying Eq. (2), we can only measure the mutation and biological functions score of genes’ pairs. To measure the importance of each gene towards triggering the disease, based on Eqs. (1-2), we quantify the driver impact quantification function $\mathbf{S}(i)$ of candidate gene i as follows:

$$\mathbf{S}(i) = \begin{cases} \max_{j \in \mathcal{N}_i} \mathbf{F}(i, j), \mathcal{N}_i \neq \emptyset \\ \mathbf{M}(i)^2/r_{max}^2, \mathcal{N}_i = \emptyset \end{cases} \quad (3)$$

where $\mathbf{F}(i, j)$ is the functional mutation score of genes i and j defined in Eq. (2), $\mathcal{N}_i$ is the set of neighbors of gene i in the network, and r_{max} is the largest average ‘interaction distance’ in the GGI and signaling networks. Eq. (3) uses the strongest functional mutation impact between the gene and its neighbors to identify possible candidate driver genes. In this way, one gene with higher mutation rate and interacting with more genes will have a higher score. This balances the mutation and functions of each gene. After this, we take $\mathbf{S}(i)$ as the *driver weight* of gene i with respect to the target disease.

2.2 Step 2: Identify cooperative driver pathways by matrix tri-factorization

Although candidate driver genes have been identified in the previous step, we still need to map these driver impacts of genes on the disease to pathways, for cooperative driver pathways discovery. Besides genes and pathways, microRNAs (miRNAs) transcribed from genes can also regulate the gene expression by the role of signal modulators and are important components of cancer pathways, and thus have direct associations with tumors. Both miRNAs and genes can act as oncogenes and tumor suppressors, and they are important components of pathways. The relation information among genes, miRNAs and pathways are relatively complete compared to other molecules. Therefore, in this step, CDPPathway fuses the inter- and intra- relations of genes, miRNAs and pathways to map driver genes to pathways and to discover cooperative driver pathways. We first construct a heterogeneous network composed of 3 types of nodes (genes, miRNAs, and pathways), three types of inter-relational networks, and three types of intra-relational networks of these nodes. The weight adjacency matrix of the heterogeneous network is formalized as follows:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{gg} & \mathbf{R}_{gm} & \mathbf{R}_{gp} \\ \mathbf{R}_{gm}^T & \mathbf{R}_{mm} & \mathbf{R}_{mp} \\ \mathbf{R}_{gp}^T & \mathbf{R}_{mp}^T & \mathbf{R}_{pp} \end{bmatrix} \quad (4)$$

where $\mathbf{R}_{gg} \in \mathbb{R}^{g \times g}$, $\mathbf{R}_{mm} \in \mathbb{R}^{m \times m}$, and $\mathbf{R}_{pp} \in \mathbb{R}^{p \times p}$ encode the intra-relations between gene, miRNA, and pathway nodes, respectively. $\mathbf{R}_{gm} \in \mathbb{R}^{g \times m}$, $\mathbf{R}_{gp} \in \mathbb{R}^{g \times p}$, and $\mathbf{R}_{mp} \in \mathbb{R}^{m \times p}$ encode the inter-relations between gene-miRNA, gene-pathway, and miRNA-pathway nodes. The superscript T is the matrix transpose operator.

To reduce the impact of passenger genes and decrease the time cost, only the gene nodes with driver weights above the specified threshold θ are included in the network. If there is a known intra-relation between two genes (i and j), then $\mathbf{R}_{gg}(i, j) = 1$, otherwise $\mathbf{R}_{gg}(i, j) = 0$. Similarly, if there is a known inter-relation between gene i and miRNA j , then $\mathbf{R}_{gm}(i, j) = 1$, otherwise $\mathbf{R}_{gm}(i, j) = 0$. The other relational matrices of the heterogeneous network are similarly specified. In addition, to import the driver impacts of genes on the disease, we update the edge weights of gene-miRNA and gene-pathway inter-relational networks based on the quantified driver weights in Step 1 as follows:

$$\mathbf{R}_{gm} = \mathbf{D}_{gg}\mathbf{R}_{gm}, \quad \mathbf{R}_{gp} = \mathbf{D}_{gg}\mathbf{R}_{gp} \quad (5)$$

where $\mathbf{D}_{gg} \in \mathbb{R}^{g \times g}$ is the driver weight matrix, $\mathbf{D}_{gg}(i, 1) = \mathbf{S}(i)$ (driver impact weight obtained in Step 1), and $\mathbf{D}_{gg}(i, j) = 0$ ($j > 1$). In this way, the impact of each candidate gene to the disease can be embedded into the inter-relational networks, and further transferred into pathways.

Nonnegative matrix factorization (NMF) and its variants can explore sub-structures of the data matrix and have been widely used to fuse multiple data types [41, 42, 43, 44]. To remedy the sparsity of the interaction data and to map genes’ influence on the carcinogenesis to pathways through the multiple relations among genes, miRNAs and pathways, we factorize the adjacency matrices of the heterogeneous network in a collaborative fashion as follows:

$$\min_{\mathbf{H}_o \geq 0} \mathcal{L}(\mathbf{H}; \mathbf{S}) = \sum_{o \in \{g, m, p\}} \|\mathbf{R}_{oo} - \mathbf{H}_o \mathbf{S}_{oo} \mathbf{H}_o^T\|_F^2 \quad (6)$$

$$- \sum_{o, o' \in \{g, m, p\}} \text{tr}(\mathbf{H}_o^T \mathbf{R}_{oo'} \mathbf{H}_{o'}) + \gamma \sum_{o \in \{g, m, p\}} \|\mathbf{H}_o\|_1^2$$

where $\|\cdot\|_F^2$ and $\text{tr}(\cdot)$ are the Frobenius norm and the matrix trace operator. $\mathbf{H}_o \in \mathbb{R}^{n_o \times k_o}$ is the low-rank representation of n_o gene (miRNA or pathway) nodes, $\mathbf{S}_{oo} \in \mathbb{R}^{k_o \times k_o}$ encodes the latent intra-relations between respective node types. $k_o \ll n_o$ is the low-rank size. $\|\mathbf{R}_{oo} - \mathbf{H}_o \mathbf{S}_{oo} \mathbf{H}_o^T\|_F^2$ is the reconstruction loss of intra-relational networks of genes (miRNAs and pathways). $\text{tr}(\mathbf{H}_o^T \mathbf{R}_{oo'} \mathbf{H}_{o'})$ is introduced to enforce the inter-relations being preserved with respect to the low-rank representations of different objects. $\|\mathbf{H}_o\|_1^2$ enforces the sparsity of $\mathbf{H}_o$ and γ is a scalar weight parameter. Eq. (6) can respect the intrinsic structure of the heterogeneous network and explore the latent relationships between genes, miRNAs, and pathways. In addition, the driver impacts embedded into the inter-relational networks can also coordinate the exploration of cooperation between pathways. The detailed optimization procedure for Eq. (6) is provided in Section 1 of the Supplementary file.

After obtaining the optimized $\mathbf{H}_o$ and $\mathbf{S}_{oo}$ in Eq. (6), we can reconstruct the pathway-pathway interaction subnetwork $\mathbf{R}_{pp}^*$ as follows:

$$\mathbf{R}_{pp}^* = \mathbf{H}_p \mathbf{S}_{pp} \mathbf{H}_p^T \quad (7)$$

Since the inter-relational gene-pathway and gene-miRNA networks import the driver weights to reconstruct the pathway-pathway interaction network, pathways with interaction and driver influence on disease have high association values in $\mathbf{R}_{pp}^*$, as our experimental results will manifest. Given this, we identify the top K pathway pairs with the highest value in $\mathbf{R}_{pp}^*$ as cooperative driver pathways.

3 Results

3.1 Datasets and parameter investigation

To investigate the effectiveness of CDPPathway, we carry out experiments on publicly available cancer data of breast cancer (BRCA) and uterine corpus endometrial carcinoma (UCEC) from TCGA [1]. Details about the

Fig. 2. Distribution of driver weights on BRCA and UCEC datasets.

multiple data sources used for the experiments are listed in Table 1. Each of the interaction weight between two genes was extracted and standardized with a value ranging from 0 to 1, by divided to the largest interaction weight among all genes. Self-loop interactions were removed to simplify the networks. Multiple miRNAs’ expressions change during tumorigenesis and development, and cascade to downstream target gene expression in the signaling pathway, and thus affect tumor progression. Given that, we assume that an miRNA and a pathway are more likely to be related if they are associated with the same gene, so we construct the miRNA-pathway association network based on the known gene-pathway and gene-miRNA associations.

We divided the original data by 7:1:2 into training data, validation data and test data, and selected the parameters using the validation data. The driver weight threshold θ is an important parameter for filtering useless genes in Step 2. We split useless genes with driver weight less than θ to reconstruct relational networks of genes, miRNAs and pathways. As a result, the reconstructed gene-gene network used in Step 2 is a sub-network of the original gene-gene interaction (GGI) network used in Step 1. We show the distribution of driver weight in Figure 2, the majority of genes’ driver weight are less than 0.01 and 0.001 for BRCA and UCEC datasets, indicating that these genes are passenger genes, which do not contribute to the carcinogenesis. To study the sensitivity of θ , based on the distribution of driver weights in Figure 2, we fix low-rank size $k_g = k_m = k_p = k_o$ as 50 and penalty coefficient γ as 100, and increase θ from 0 to 0.1 in BRCA dataset and from 0 to 0.01 in UCEC dataset. Figure 3 reports the the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC) under different input values of θ in reconstructing pathway-pathway associations.

Both the AUROC value and AUPRC value vary no more than 0.2 as the increase of θ , and reach to the highest when $\theta \approx 0.07$ on the BRCA dataset and $\theta \approx 0.006$ on the UCEC dataset. Particularly, when $\theta = 0$, both the AUROC and AUPRC are the lowest, indicating that filtering passenger genes is helpful for reconstructing the pathway-pathway interaction network. In practice, we also tested with a larger threshold, like 0.5, 1 and 5, the AUROC and AUPRC values drop to lower than when $\theta = 0$, since important driver genes are excluded with such a large threshold. Given these observations, we adopt $\theta = 0.07$ and $\theta = 0.006$ for experiments on BRCA and UCEC datasets, respectively.

The low-rank size k_o and penalty coefficient γ are important parameters for reconstructing pathway-pathway interaction network, and tuning them altogether is time consuming. For simplicity, we simply fix $k_g = k_m = k_p = k_o$, and then study how k_o and γ affect the

reconstruction of pathway-pathway interaction network. We increase k_o from 10 to 110 and γ from 10^{-4} to 10^6 on BRCA and UCEC datasets. Figure 4 reports the AUROC and AUPRC under different input values of k_o and γ .

Both the AUROC and AUPRC increase as the increase of k_o and reach to the plateau when $k_o \geq 40$ on the two datasets. That is because the intra-relations cannot be preserved when the size (k_o) of low-rank representation $\mathbf{H}_o$ is too small. Similarly, both the AUROC and AUPRC increase as the rise of γ and reach to a plateau when $\gamma \geq 10$ on both datasets. The values of AUROC and AUPRC manifest a significant reduce when $\gamma < 1$. That is because $\gamma \sum_{o \in \{g, m, p\}} |\mathbf{H}_o|_1^2$ encourages the sparsity of $\mathbf{H}_o$ and thus to reduce the false positive reconstructed edges between pathway nodes. A too low γ cannot enforce the sparsity.

Based on the above analysis, we adopt $k_o = 90$ and $k_o = 100$, $\gamma = 10^3$ and $\gamma = 10^3$ for experiments on BRCA and UCEC datasets, respectively. Based on these selected parameters, CDPathway again performs well on the test data with an AUROC of 0.8952 and 0.9041, an AUPRC of 0.5714 and 0.5846 on the BRCA and UCEC dataset, respectively.

3.2 The driver weight quantification function can effectively identify driver genes

Known driver genes can be identified by our driver weight quantification function with statistic significance. To study the effectiveness of identifying known driver genes (reported in [53]), we specify the number of identified candidate driver genes in Step 1 from one to the number of genes in the mutation profiles (13965 in BRCA and 18421 in UCEC), and plot the ROC and PR curves in Figure 5.

For both BRCA and UCEC datasets, the AUROC values are over 0.9, and AUPRC values are over 0.2. The reason why the AUPRC is relatively low compared to the AUROC is due to that CDPathway identifies genes which drive the carcinogenesis of the target cancer, while the known genes are drivers for various diseases. Another cause is that AUPRC is more sensitive to the class-imbalance task and identifying known driver gene in essence is a class-imbalance task. As can be seen, the Precision is very high at the beginning, which indicates that top ranked genes are important driver genes. The genes with weights less than θ should be excluded from the heterogeneous network, and they drag down the Precision for a high Recall and cause a low AUPRC value. Overall, CDPathway can effectively identify known driver genes.

Fig. 3. AUROC and AUPRC vs. θ .

Fig. 4. The AUROC and AUPRC of CDPathway under different low-rank size k_o and sparsity coefficient γ .

Table 1. Details of collected datasets.

Data type	Size(BRCA)	Size(UCEC)	Direction	Weight	Sources	Version or Date
mutation profile	13965 $\times$ 776	18421 $\times$ 248	No	No	TCGA [1]	2018.11
GGI	16907 $\times$ 16907	16907 $\times$ 16907	Undirected	Weighted	STRING [59]	2017.05
signaling network	12456 $\times$ 12456	12456 $\times$ 12456	Directed	Binary	SignalLink [46], SPIKE [47, 48] Pathway Commons [49]	2018.11
gene-gene ($\mathbf{R}_{gg}$)	1631 $\times$ 1631	2010 $\times$ 2010	Undirected	Binary	reconstructed	2019.05
gene-miRNA ($\mathbf{R}_{gm}$)	1631 $\times$ 559	2010 $\times$ 559	Undirected	Weighted	miRBase [50]	2019.03
gene-pathway ($\mathbf{R}_{gp}$)	1631 $\times$ 212	2010 $\times$ 212	Undirected	Weighted	PID [51]	2018.01
miRNA-miRNA ($\mathbf{R}_{mm}$)	559 $\times$ 559	559 $\times$ 559	Undirected	Binary	Cui's lab [52]	2018.01
miRNA-pathway ($\mathbf{R}_{mp}$)	559 $\times$ 212	559 $\times$ 559	Undirected	Binary	reconstructed	2019.05
pathway-pathway ($\mathbf{R}_{pp}$)	212 $\times$ 212	212 $\times$ 212	Undirected	Binary	PID [51]	2018.01

Fig. 5. Receiver operating characteristics (ROC) and precision recall (PR) curves of identifying known driver genes for BRCA and UCEC datasets. The corresponding AUROC and AUPRC values are also included in the legend. CDPATHWAY-noSig: CDPATHWAY without the signaling network. CDPATHWAY-noGGI: CDPATHWAY without the GGI network. CDPATHWAY-noNetwork: CDPATHWAY without the signaling and GGI networks.

In order to investigate the contribution of GGI and signaling networks, we also compare CDPATHWAY with CDPATHWAY without the GGI or the signaling networks. When a single functional network is disregarded, the AUROC and AUPRC values are reduced by 1.7% and 16.7% on BRCA, and by 1.8% and 13.1% on UCEC, respectively. When both networks are removed, the AUROC (AUPRC) value drops sharply by 15.1% (62.8%) and 16.3% (53.6%) on BRCA and UCEC datasets, respectively. Therefore, we can conclude that both prior knowledge networks contribute to the identification of driver genes, and disregarding any of the two will compromise the performance. In addition, we also study how CDPATHWAY perform when the functional networks are not complete. We randomly remove 5%-50% edges in the GGI or signaling network and record the AUROC and AUPRC values in Fig. 6. As can be seen, the AUROC and AUPRC manifest a decrease as more edges removed. Generally, CDPATHWAY relies to the completeness of functional networks, but the AUROC (AUPRC) drops less than 10% (25%) when even 50% of edges in the GGI or signaling network are removed. That is because we also integrate the somatic mutation data in Step 1, and the mutation rate of genes (i.e., coverage) are helpful to identify potential driver genes. Therefore, we can infer that the driver weights of genes are not so dependent on the completeness of the functional networks.

After filtering useless genes, we further apply a hypergeometric test on the identified candidate driver genes of the two real datasets to study the statistical significance of the candidate driver genes. The p -values of the identified known driver genes are $9.68E-62$ and $1.67E-54$ for BRCA and UCEC datasets, respectively. Therefore, we can conclude that CDPATHWAY can identify known driver genes with statistical significance. In addition, we select the top 30 candidate driver genes in Step 1, and

do functional and pathway enrichment analysis using the widely-used DAVID toolkit [54]. As Table 2 shows, there are 17 and 14 known driver genes among the 30 candidate driver genes for BRCA and UCEC datasets, respectively. Besides the known driver genes, CDPATHWAY can also identify several novel driver genes. For the BRCA dataset, it is highlighted that FOXA1 (Forkhead box protein A1) and GATA3 (GATA binding protein 3) co-regulate the expression of genes, which is essential for the luminal mammary epithelial cell development [55, 56]. Another example is that mutations in MAP2K4 produce perturbations similar to MAP3K1 loss [57]. For UCEC dataset, TP53 (tumor protein p53) is involved with DNA repair, growth arrest, and apoptosis. In particular, mutations in TP53 can lead to uncontrolled proliferation and invasive growth. On the other hand, DCC (Deleted in Colorectal Carcinoma) is reported to have an anti-metastatic role [58], meaning that it may only contribute to cancer in the context of a pre-existing condition. We conjecture that the mutations in DCC may contribute to cancer progression for patients with defective mismatch repair and/or impaired TP53 functionality. A network based analysis also provides supporting evidence of TTN mutations as a disease marker [60].

Table 2 also shows some enriched function terms. There are several overlapping important biological processes enriched in both BRCA and UCEC datasets, including transcriptional activator activity, central carbon metabolism in cancer, apoptosis, etc. The candidate driver genes are also enriched in the target disease and cancer related pathways. These results demonstrate that CDPATHWAY can effectively identify candidate driver genes engaging in the cancer related pathways and biological processes.

Fig. 6. AUROC and AUPRC values of identifying known driver genes with different proportions of removed interactions in the GGI and signaling networks. CDPPathway-rGGI(Sig): CDPPathway uses only the GGI(signaling) network with removed interactions. CDPPathway-GGI(Sig): CDPPathway using only the GGI(signaling) network.

Table 2. Candidate driver genes identified by CDPPathway. The known driver genes and enriched biological processes in BRCA and UCEC datasets are highlighted in boldface font.

Dataset	Genes	GO/KEGG/Disease enrichment
BRCA	TP53 PIK3CA CDH1 PTEN BRCA2 MAP3K1 AKT1 CTCF PIK3R1 RUNX1 GATA3 FOXA1 NCOA3 NCOR1 DMD MAP2K4 ERBB2 NF1 ARID1A LRP2 ATM KRAS ACTN2 FLT4 MTOR RELN LRBA ARHGEF1 CASP8 DCC	Breast cancer; Pathways in cancer; Central carbon metabolism in cancer; Apoptosis; MicroRNAs in cancer; Transcriptional activator activity; RNA polymerase II transcription regulatory region sequence-specific binding
UCEC	PTEN PIK3CA CTNNB1 TP53 PIK3R1 ARID1A CTCF KRAS FGFR2 FBXW7 PPP2R1A TTN SGK1 DMD ESR1 SPOP CCND1 TNS1 MTOR FAT1 LRP6 ERBB3 CUX1 MED12 DCC APC SOS1 NFE2L2 CTNND1 ZMYM2	Endometrial cancer; Pathways in cancer; Central carbon metabolism in cancer; MicroRNAs in cancer; Negative regulation of transcription from RNA polymerase II promoter

3.3 Collaborative matrix factorization can effectively reconstruct the pathway-pathway interaction network

To investigate the effectiveness of collaborate matrix factorization in Step 2, we randomly divide the original pathway-pathway interaction network into five folds for cross validation. Next, we plot the ROC and PR curves of CDPPathway in Figure 7. We run five fold cross validation for

ten independent rounds, and report the average results. We can see that CDPPathway achieves high AUROC values of 0.9419 and 0.9318, AUPRC values of 0.7838 and 0.7707 on BRCA and UCEC datasets, respectively. The high AUROC and AUPRC values indicate that CDPPathway can accurately reconstruct pathway-pathway associations.

Fig. 7. ROC and PR curves of reconstructing the pathway-pathway interaction network for BRCA and UCEC datasets. CDPATHWAY-noSig: CDPATHWAY without the signaling network. CDPATHWAY-noGGI: CDPATHWAY without the GGI network. CDPATHWAY-noNetwork: CDPATHWAY without the signaling and GGI networks. CDPATHWAY-noWeight: CDPATHWAY without the weights of driver genes.

To investigate the contribution of the quantified driver gene weights obtained in Step 1, we test the collaborate matrix factorization without importing driver weights. As shown in Figure 7, both AUROC and AUPRC drop by 12.6% and 8.3%, 27.2% and 23.4% on BRCA and UCEC datasets, respectively, indicating that the imported weights are helpful in reconstructing the pathway-pathway interaction network. In addition, we also test CDPATHWAY without prior functional networks. Without any functional network, both the AUROC and AUPRC values drop more seriously. The integration of the signaling and GGI networks can jointly improve the performance. In addition, the GGI and signaling networks show equal importance (no more than 3% difference) for the reconstruction of the pathway-pathway interaction network. We can conclude that the GGI and signaling networks complement each other and both contribute to the reconstruction of the pathway-pathway interaction network.

3.4 CDPATHWAY can identify significant cooperative driver pathways

We select 10 pathway pairs with top 10 highest values in the reconstructed pathway-pathway interaction network as cooperative driver pathways. As shown in Table 3, by checking the member genes of these pathways, we find that each pathway contains at least one known driver gene, which indicates

that all pathways play important roles in carcinogenesis by undertaking the biological activities of the driver genes. In addition, these pathways also confirm that the driver effects of genes on the disease are correctly transferred to pathways by the collaborative matrix factorization on the heterogeneous network in Step 2.

We further check their cooperation and find that all the pathway pairs are associated in the collected pathway-pathway interaction network. We visualize pathway pair 1 of BRCA dataset in Figure 8 as an example. Other examples are provided in the Supplementary file. We observe that the pathway pair 1 interacts with each other to undertake biological activities. These two pathways also share common known driver genes: PIK3CA, PIK3R1, and KRAS. These genes have interactions and may cooperate with each other to drive the target disease. The shared gene NCOA3 does not have evident functional edges with any other gene, but it was predicted to have connections with the driver gene EGFR [59], which also supports the capability of CDPATHWAY in identifying cooperative driver pathways.

There are 5 known pan-cancer related pathways identified by CDPATHWAY, including Class I PI3K signaling events mediated by Akt, mTOR signaling pathway, TGF- β receptor signaling, Canonical Wnt signaling pathway, and Notch signaling pathway [61]. Breast cancer studies have shown that the abnormality of Internalization of ErbB1, Canonical Wnt signaling pathway, Alternative NF-kappaB pathway, Class

Fig. 8. Pathway pair 1 of BRCA dataset in the GGI network from String [59]. The rose red lines indicate interactions from curated databases and blue lines indicate interactions have been experimentally determined. Genes in the ‘Internalization of ErbB1’ pathway are circled, and genes in ‘Trk receptor signaling mediated by PI3K and PLC-gamma’ pathway are enclosed by rectangles. These two pathways share common known driver genes of PIK3CA, PIK3R1, and KRAS.

I PI3K signaling events mediated by Akt, Notch signaling pathway and C-MYC pathway promote the carcinogenesis [62, 63, 64, 65, 66, 67]. For UCEC, Trk receptor signaling mediated by PI3K and PLC-gamma, TGF-beta receptor signaling, FOXM1 transcription factor network and mTOR signaling pathway are reported to have tight associations with the target disease [68, 69, 70].

We can also find that identified driver pathway pairs play important roles in the life activities related to the initiation of cancer. For example, in pathway pair 3 (Figure S12 of Supplementary file) on the UCEC dataset, ATM pathway provides the crucial link between DNA damage, cell cycle progression and cell death by first sensing double stranded DNA breaks, and subsequently phosphorylating and activating other downstream proteins functioning in DNA damage repair, cell cycle arrest and apoptotic pathways [71]. Therefore, the identified pathways’ pairs cooperate with each other and play important roles in biological activities. It’s quite possible that their cooperation drives the occurrence of the target disease.

3.5 Comparison with other methods

We compare CDPATHWAY against six related and competitive methods, including two individual driver pathways identification methods: Dendrix [13] and BLP [14]; and four multiple driver pathways identification methods: CoMDP [27], Multi-Dendrix [26], BeCo-WithMEFun [28], and CoDP [29]. We discussed these approaches in the Introduction. We set the parameter ranges of the comparing methods as described in the original papers, and select pathways that contain the most known driver genes with statistic significance. Since these methods, except CoDP, do not identify exact driver pathways, we perform a comparison at the gene level, by mapping pathways identified by CDPATHWAY and CoDP to genes

using known gene-pathway association networks. On the other hand, we compare CDPATHWAY with these methods at the pathway level by doing a KEGG enrichment test for the driver gene set identified by comparing methods via DAVID [54].

As shown in Table 4, CDPATHWAY outperforms other methods by identifying on average twice more known driver genes. **Biologically more genes make less sense, but CDPATHWAY uncovers more driver genes instead of passenger genes than other methods. In fact, CDPATHWAY uncovers how driver pathways cooperatively trigger cancer by including these driver genes to affect important biological activities. Therefore, CDPATHWAY provides driver pathways with diversity.** As shown in Table 4, the known driver genes are identified with statistical significance (Hypergeometric test P-value <0.05), and has better statistical significance and F1-Score than other comparing methods, except getting a similar F1-Score with CoMDP on the UCEC dataset. Therefore CDPATHWAY does not identify more known driver genes by only identifying more pathways, it balances the Precision and Recall of identifying cooperative driver pathways. In contrast, the comparing methods identify limited driver genes. As a result, they also provide limited diversity of driver pathways in different datasets. These results indicate that CDPATHWAY can better capture the heterogeneity of different datasets than comparing methods.

We also compare CDPATHWAY with comparing methods on identifying known driver genes. We fix the number of identified driver genes from 1 to the number of genes in the somatic mutation profile, and show the ROC and PR curves in Fig. 9. Since BeCo-WithMEFun cannot get a solution when the number of identified genes is larger than 40, its ROC and PR curves cannot be obtained. As shown in Fig. 9, CDPATHWAY outperforms the comparing methods with respect to AUROC and AUPRC by at least 15%, since it makes use of somatic mutation data and prior knowledge

Table 3. Cooperative driver pathways identified by CDPPathway.

No.	BRCA	UCEC
1	Internalization of ErbB1, Trk receptor signaling mediated by PI3K and PLC-gamma	JNK signaling in the CD4 TCR pathway, HIV-1 Nef Negative effector of Fas and TNF-alpha
2	Degradation of beta catenin, Canonical Wnt signaling pathway	Internalization of ErbB1, Trk receptor signaling mediated by PI3K and PLC-gamma
3	Ceramide signaling pathway, EPO signaling pathway	ATM pathway, Reelin signaling pathway
4	IL2 signaling events mediated by STAT5, Osteopontin-mediated events	Canonical Wnt signaling pathway, BCR signaling pathway
5	Alternative NF-kappaB pathway, HIF-1-alpha transcription factor network	TGF-beta receptor signaling, Ceramide signaling pathway
6	p53 pathway, Class I PI3K signaling events mediated by Akt	FOXA1 transcription factor network, Regulation of Androgen receptor activity
7	Aurora A signaling, Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	mTOR signaling pathway, IL2 signaling events mediated by STAT5
8	Notch signaling pathway, Integrin family cell surface interactions	TNF receptor signaling pathway, FOXO1 transcription factor network
9	Aurora C signaling, C-MYC pathway	Netrin-mediated signaling events, Alternative NF-kappaB pathway
10	IL8- and CXCR1-mediated signaling events, Integrins in angiogenesis	Insulin-mediated glucose transport, Direct p53 effectors

Table 4. Known driver gene sets of BRCA and of UCEC identified by CDPPathway and other comparing methods.

Method	Driver Genes of BRCA	Enrichment	F1-Score	Driver Genes of UCEC	Enrichment	F1-Score
CDPathway	PIK3CA,PIK3R1,KRAS,EGFR, BRAF,AKT1,APC,AXIN1, CTNNB1,MAP3K1,CASP8,RB1,JAK2, PTPN11,JAK1,EP300, CREBBP,SMAD4,GATA2,MET, TP53,CDH1,TNFAIP3,FBXW7	6.28E-26	0.3912	MAP3K1,CASP8,PIK3CA,PIK3R1, KRAS,NRAS,EGFR,BRAF, AKT1,ATM,GATA3,JAK2, PTPN11,JAK1,TP53, CTNNB1, APC,CDH1,MET,EP300, CREBBP	1.52E-23	0.3105
CoDP	FGFR2,FGFR3,MET,CDKN2A,EP300,TP53,MAP2K1,PIK3CA, PIK3R1,CTNNB1,AKT1,PTPN11,BRAF,CASP8,MAP3K1	5.27E-5	0.1222	NRAS,MAP2K1,B2M,HRAS,KRAS,BRAF,JAK2,SOC1, CTNNB1,KLF4,CDKN2A,MED12,CDH1,AR,EP300,APC, SMARCA4,AKT1,PIK3CA	2.32E-10	0.1935
BeCo-WithMEFun	NCOA3 TP53	3.82E-03	0.1111	ARID1A TP53	2.70E-03	0.1176
Dendrix	CDH1,TP53	2.56E-03	0.1250	PIK3CA,PIK3R1	1.83E-03	0.1333
BLP	TP53, PIK3CA, GATA3	5.19E-05	0.1764	TP53, PTEN	1.83E-03	0.1333
Multi-Dendrix	CDH1, TP53, GATA3, PIK3CA	6.85E-07	0.2353	PIK3CA, PIK3R1	1.83E-03	0.1333
CoMDP	MAP3K1, TP53, CDH1, GATA3, MLL3, PIK3CA	3.81E-11	0.3529	PIK3R1, FBXW7, PIK3CA, PPP2R1A, TP53, CTNNB1, PTEN	4.97E-14	0.3889

of gene-gene interaction network to identify potential driver genes. We observe that multiple driver pathway identification methods outperform individual driver pathway identification methods, since the former methods uncover driver pathway with diversity, while the latter ones only uncover very limited driver genes. In addition, the PR curves of these comparing methods drop sharply with the increased number of identified genes. That is because these individual driver pathway identification methods can only identify one or two gene sets (i.e., pathway), and they cannot provide meaningful results when identifying more than 10 genes within only one or two pathways.

CDPathway can also uncover novel driver genes, which play important roles in important biological activities. For example, we obtain EGF (Epidermal growth factor) in the BRCA dataset, and EGF has tight associations with the survival rate of breast cancer patients [72]. The reason why CDPPathway outperforms other methods is that CDPPathway integrates both mutation and GGI data, which help reflecting the importance of genes, and map weights of driver genes to pathways. As a result, its identified cooperative driver pathways contain more known (potential) driver genes. These results suggest that instead of considering individual genes, transferring the driver impacts of genes to pathways through their associations can better explore how pathways cooperatively drive cancer.

As shown in Tables 5 and 3, pathways identified by these methods also take responsibility for disease related activities. Overall, 40% of the pathways identified by CDPPathway and by the comparing methods are overlapped, which support CDPPathway’s accuracy in identifying disease related pathways. For example, for UCEC, the ‘mTOR Signaling Pathway’ was identified as the driver pathway by Dendrix, Multi-Dendrix, and

CoMDP, and also by CDPPathway. The ‘mTOR Signaling Pathway’ is ubiquitously expressed in cells and is a therapeutic target for the cancer treatment arsenal. Our CDPPathway can identify more significant driver pathways by 150% than the individual driver identification methods, and it also can discover more driver pathways with cooperation by 60% than cooperative driver pathway identification methods. Particularly, only 60% of the cooperation among pathways identified by CoDP (most effective method among the comparing methods) is supported by the pathway-pathway interaction network, while that of CDPPathway is 100%. This observation indicates that CDPPathway is more effective in identifying pathway pairs with both cooperation and functional significance. CDPPathway can accurately identify important genes in Step 1 and Step 2, can retain the driver effects of genes on interacting pathways, and thus more accurately identify cooperative driver genes. Because of the driver genes, these identified pathways not only have interaction with each other, but also cooperatively carry out disease related activities.

4 Discussion

Cooperative driver pathways discovery helps understanding how multiple pathways function cooperatively in cancer initiation and progression, and provides more precise therapy to patients. For this discovery, we proposed a two-stage based approach called CDPPathway. CDPPathway first uses a driver impact quantification function to identify candidate driver genes. Next, CDPPathway collaboratively factorizes the adjacency matrices of a heterogeneous network of genes, miRNAs, and pathway

Fig. 9. Comparison of ROC and PR curves of identifying known driver genes for BRCA and UCEC datasets.

Table 5. Cooperative driver pathways identified by other comparing methods.

Method	BRCA	UCEC
CoDP	Beta1 integrin cell surface interactions, Syndecan-1-mediated signaling events; Validated transcriptional targets of AP1 family members Fra1 and Fra2, Signaling mediated by p38-alpha and p38-beta; Signaling events mediated by VEGFR1 and VEGFR2, Ceramide signaling pathway; Direct p53 effectors, Validated nuclear estrogen receptor alpha network;	Downstream signaling in naive CD8 T cells, IL12-mediated signaling events; Regulation of RAC1 activity, Regulation of RhoA activity; Regulation of nuclear beta catenin signaling and target gene transcription, IL4-mediated signaling events; TCR signaling in naive CD4 T cells, Validated targets of C-MYC transcriptional repression;
BeCo-WithMEFun	p53 pathway, Class I PI3K signaling events mediated by Akt;	B cell receptor signaling pathway, T cell receptor signaling pathway;
Dendrix	Thyroid cancer	Trka Receptor Signaling Pathway, mTOR Signaling Pathway
BLP	Small cell lung cancer	p53 signaling pathway
Multi-Dendrix	Thyroid cancer,p53 pathway;	mTOR signaling pathway, IL2 signaling events mediated by STAT5;
CoMDP	Thyroid cancer,p53 pathway	ErbB signaling pathway, mTOR signaling pathway; Apoptosis, B cell receptor signaling pathway;

to fuse multi-relational data. CDPPathway then reconstructs the pathway-pathway interaction network and identifies the pathway pairs with the highest reconstruction values as the cooperative driver pathways.

Experimental results on public breast, endometrial and ovarian cancer datasets show that CDPPathway not only uncovers known driver genes with better statistical significance than competitive methods, but also discovers more novel driver genes. We observe that integrating mutation data and prior functional networks can better differentiate driver genes from passenger genes, because CDPPathway uncovers genes with high mutation rate, which are hub nodes in the functional networks. The collaborative

matrix factorization correctly transfers these driver genes to pathways, and thus the cooperative driver pathways identified by CDPPathway contain more known driver genes than those of existing methods. Additionally, less known and less frequently altered genes with well characterized cancer drivers, suggest a mechanism of action. For example, on BRCA dataset, FOXA1 and GATA3 co-regulate the expression of genes, which are essential for the luminal mammary epithelial cell development. On the pathway level, CDPPathway can identify more potential biological cooperation among pathways. On both BRCA and UCEC datasets, each identified pathway contains at least one known driver gene, and we found

the cooperation of Internalization of ErbB1 and Trk receptor signaling mediated by PI3K and PLC-gamma may be important to the occurrence of disease. We can infer that because of the tight associations within and between genes, miRNAs and pathways, transferring the driver impact from genes to pathways through miRNAs helps to uncover cooperative driver pathways.

We plan to extend this work to other biological networks. Given that there are many other types of molecules important to biological activities, such as lncRNAs, genes and drugs. Our method can be further applied on multi-type related biological networks to uncover more diseased related molecules. The challenge of extending our method is to build such relational biological networks and to choose a type of molecules with trigger influence on the disease.

Key Points

- Pathways undertake important biological activities and are robust by providing reliable gene regulatory relationships. Biological pathways have been applied to explore the pathology involved with cancer occurrence, diagnosis and prognosis. In practice, multiple pathways cooperatively drive the occurrence and progression of cancer, but scanty computational models can identify cooperative driver pathways.
- We propose a novel model CDPATHWAY to identify cooperative driver pathways. CDPATHWAY combines the mutation data and prior knowledge of genes (i.e., gene-gene interaction network) to capture the influences of genes on the occurrence of cancers. CDPATHWAY further exploits these influences to explore cooperative driver pathways via tightly associated biological networks of genes, miRNAs and pathways.
- CDPATHWAY shows a good prognostic performance in breast, endometrial and ovarian cancers. CDPATHWAY can identify driver pathways with significance and diversity. It not only identifies known driver genes already explored by existing methods, but also finds out novel driver genes. In addition, CDPATHWAY uncovers the cooperations between driver pathways, which better reflects how biological activities trigger cancer than existing methods.

Funding

This work is supported by Natural Science Foundation of China (61873214 and 61872300), Fundamental Research Funds for the Central Universities (XDJK2019B024), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228).

References

1. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011; **474**: 609-15.
2. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* 2010; **464**: 993-8.
3. Beroukhi R, Getz G, Nghiemphu L, *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Natl Acad Sci* 2007; **104**: 20007-12.
4. Ding L, Getz G, Wheeler DA, *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 2008; **455**: 1069.
5. Torkamani A, Schork NJ. Identification of rare cancer driver mutations by network reconstruction. *Genome Res* 2009; **19**: 1570-8.
6. Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61.
7. Bashashati A, Haffari G, Ding J, *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* 2012; **13**: R124.
8. Yang Q, Wang S, Dai E, *et al.* Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Brief Bioinform*, 2017, **20**: 168-77.
9. Karp PD, Latendresse M, Paley SM, *et al.* Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform*, 2015, **17**: 877-90.
10. Li F, Wu T, Xu Y, *et al.* A comprehensive overview of oncogenic pathways in human cancer. *Brief Bioinform*, 2019.
11. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nature Med* 2004; **10**: 789.
12. Yeang C H, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J* 2008; **22**: 2605-22.
13. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res* 2012; **22**: 375-85.
14. Zhao J, Zhang S, Wu LY, *et al.* Efficient methods for identifying mutated driver pathways in cancer. *Bioinform* 2012; **28**: 2940-47.
15. Zhang J, Zhang S. Discovery of cancer common and specific driver gene sets. *Nucleic Acids Res* 2017; **45**: e86.
16. Wendl MC, Wallis JW, Lin L, *et al.* PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinform* 2011; **27**: 1595-1602.
17. Dees ND, Zhang Q, Kandoth C, *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* 2012; **22**: 1589-98.
18. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 2011; **18**: 507-22.
19. Cerami E, Demir E, Schultz N, *et al.* Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, 2010; **5**: e8918.
20. Hua X, Xu H, Yang Y, *et al.* DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am J Human Genet* 2013; **93**: 439-51.
21. Jia P, Zhao Z. VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput Biol* 2014; **10**: e1003460.
22. Leiserson MDM, Vandin F, Wu HT, *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015; **47**: 106.
23. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*, 2011; **144**: 646-74.
24. Cui Q, Ma Y, Jaramillo M, *et al.* A map of human cancer signaling. *Molecular Sys Biol* 2007; **3**.
25. Zhao Y, Ma J, Fan Y, *et al.* TGF- β transactivates EGFR and facilitates breast cancer migration and invasion through canonical Smad3 and ERK/Sp1 signaling pathways. *Mol Oncol* 2018; **12**: 305-21.
26. Leiserson MDM, Blokh D, Sharan R, *et al.* Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* 2013; **9**: e1003054.
27. Zhang J, Wu LY, Zhang XS, *et al.* Discovery of co-occurring driver pathways in cancer. *BMC Bioinform* 2014; **15**: 271.
28. Dao P, Kim YA, Wojtowicz D, *et al.* BeWith: A Between-Within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. *PLoS Comput Biol* 2017; **13**: e1005695.
29. Yang Z, Yu G, Guo M, *et al.* CoDP: Cooperative Driver Pathways Discovery With Matrix Factorization and Tri-Random Walk. *IEEE Access* 2019; **7**: 77738-49.
30. Gu Y, Wang H, Qin Y, *et al.* Network analysis of genomic alteration profiles reveals co-altered functional modules and driver genes for glioblastoma. *Mol Biosyst* 2013; **9**: 467-77.

31. Wood LD, Parsons DW, Jones S, *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* 2007; **318**: 1108-13.
32. Babur O, Gonen M, Aksoy BA, *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol* 2015; **16**: 45.
33. Hou Y, Gao B, Li G, *et al.* MaxMIF: A New Method for Identifying Cancer Driver Genes through Effective Data Integration. *Adv Sci* 2018; **5**: 1800640.
34. Simini F, Gonzalez MC, Maritan A, *et al.* A universal model for mobility and migration patterns. *Nature* 2012; **484**: 96.
35. Cheng F, Liu C, Lin CC, *et al.* A gene gravity model for the evolution of cancer genomes: a study of 3,000 cancer genomes across 9 cancer types. *PLoS Comput Biol* 2015. **11**: e1004497.
36. Teng XC, Dayhoff BM, Cheng WC, *et al.* Genome-wide consequences of deleting any single gene. *Molecular cell* 2013. **52**: 485-94.
37. Emerling BM, Hurov JB, Poulogiannis G, *et al.* Depletion of a putatively druggable class of phosphatidylinositol kinases inhibits growth of p53-null tumors. *Cell* 2013. **155**: 844-57.
38. Liu Y, Zhang X, Han C, Wan G, *et al.* TP53 loss creates therapeutic vulnerability in colorectal cancer. *Nature* 2015. **520**: 697-701.
39. Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. *Cell* 2012; **148**: 1172-87.
40. Pian C, Zhang G, Gao L, *et al.* miR+ Pathway: the integration and visualization of miRNA and KEGG pathways. *Brief Bioinform*, 2019.
41. Zitnik M, Zupan B. Data fusion by matrix factorization. *IEEE Trans Pattern Anal Mach Intell* 2014; **37**: 41-53.
42. Wang H, Zheng C, Zhao X. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. *Bioinform* 2014; **31**: 572-80.
43. Fu G, Wang J, Domeniconi C, Yu G. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinform* 2018; **34**: 1529-1537.
44. Yu G, Wang Y, Wang J, Fu G, Guo M, Domeniconi C. Weighted matrix factorization based data fusion for predicting lncRNA-disease associations. *IEEE International Conference on Bioinformatics and Biomedicine*, 2018, pp. 572-577.
45. Chatr-Aryamontri A, Oughtred R, Boucher L, *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 2017; **45**: D369-79.
46. Fazekas D, Koltai M, Turei D, *et al.* SignaLink 2-a signaling pathway resource with multi-layered regulatory networks. *BMC Sys Biol* 2013; **7**: 7.
47. Paz A, Brownstein Z, Ber Y, *et al.* SPIKE: a database of highly curated human signaling pathways. *Nucleic Acids Res* 2010; **39**: D793-99.
48. Klingström T, Plewczynski D. Protein-protein interaction and pathway databases, a graphical review. *Brief Bioinform*, 2010; **12**: 702-13.
49. Cerami EG, Gross BE, Demir E, *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* 2010; **39**: D685-90.
50. Griffiths-Jones S, Saini HK, van Dongen S, *et al.* miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2007; **36**: D154-58.
51. Schaefer C F, Anthony K, Krupa S, *et al.* PID: the pathway interaction database. *Nucleic Acids Res* 2008; **37**: D674-D79.
52. Wang D, Wang J, Lu M, *et al.* Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatic* 2010; **26**: 1644-50.
53. Vogelstein B, Papadopoulos N, Velculescu VE, *et al.* Cancer genome landscapes. *Science* 2013; **339**: 1546-58.
54. Dennis G, Sherman BT, Hosack DA, *et al.* DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003; **4**: R60.
55. Theodorou V, Stark R, Menon S, *et al.* GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res* 2013; **23**: 12-22.
56. Bernardo GM, Bebek G, Ginther CL, *et al.* FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene* 2013; **32**: 554.
57. Ellis MJ, Ding L, Shen D, *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* 2012; **486**: 353.
58. Arakawa H. Netrin-1 and its receptors in tumorigenesis. *Nat Rev Cancer* 2004; **4**: 978.
59. Franceschini A, Szklarczyk D, Frankild S, *et al.* STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2012; **41**: D808-15.
60. Hofree M, Shen JP, Carter H, *et al.* Network-based stratification of tumor mutations. *Nat Methods* 2013; **10**: 1108.
61. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**: 27-30.
62. Slamon DJ. The FUTURE of ErbB-1 and ErbB-2 pathway inhibition in breast cancer: targeting multiple receptors. *Oncologist*, 2004; **9**: 1-3.
63. Howe LR, Brown AMC. Wnt signaling and breast cancer. *Cancer Biol Ther* 2004; **3**: 36-41.
64. Zhou J, Zhang H, Gu P, *et al.* NF- κ B pathway inhibitors preferentially inhibit breast cancer stem-like cells. *Breast Cancer Res Treat* 2008; **111**: 419-27.
65. Vara JAF, Casado E, Castro J, *et al.* PI3K/Akt signalling pathway and cancer. *Cancer Ther Rev* 2004; **30**: 193-204.
66. Stylianou S, Clarke RB, Brennan K. Aberrant activation of notch signaling in human breast cancer. *Cancer Res* 2006; **66**: 1517-25.
67. Liao DJ, Dickson RB. c-Myc in breast cancer. *Endocr Relat Cancer* 2000; **7**: 143-64.
68. Piestrzeniewicz UD, Brys M, Semczuk A, *et al.* TGF- β signaling is disrupted in endometrioid-type endometrial carcinomas. *Gynecol Oncol* 2004; **95**: 173-180.
69. Kang S, Dong SM, Kim BR, *et al.* Thioridazine induces apoptosis by targeting the PI3K/Akt/mTOR pathway in cervical and endometrial cancer cells. *Apoptosis* 2012; **17**: 989-97.
70. Chan DW, Yu SYM, Chiu PM, *et al.* Over-expression of FOXM1 transcription factor is associated with cervical cancer progression and pathogenesis. *J Pathol* 2008; **215**: 245-52.
71. Kitagawa R, Kastan MB. The ATM-dependent DNA damage signaling pathway. *Cold Spring Harbor Symposia on Quantitative Biology* 2005; **70**: 99-109.
72. Mizukami Y, Nonomura A, Noguchi M, *et al.* Immunohistochemical study of oncogene product ras p21, c-myc and growth factor EGF in breast carcinomas. *Anticancer Res* 1991; **11**: 1485-94.