UC Irvine

UC Irvine Previously Published Works

Title

A global biomolecular network alignment method based on network flow model

Permalink

<https://escholarship.org/uc/item/5166b3r2>

ISBN

9781509030491

Authors

Xie, J
Li, J
Wang, J
et al.

Publication Date

2017-12-15

DOI

10.1109/BIBM.2017.8217924

Peer reviewed

KERNEL DIFFERENTIAL SUBGRAPH REVEALS DYNAMIC CHANGES IN BIOMOLECULAR NETWORKS

Jiang Xie*

*School of Computer Engineering and Science, Shanghai University
99 Shang Da Road, Shanghai 200444, China
jiangx@shu.edu.cn*

Dongfang Lu

*School of Computer Engineering and Science, Shanghai University
99 Shang Da Road, Shanghai 200444, China
ludongfang@shu.edu.cn*

Jiaxin Li

*School of Computer Engineering and Science, Shanghai University
99 Shang Da Road, Shanghai 200444, China
jiaxinlee@i.shu.edu.cn*

Jiao Wang*

*Laboratory of Molecular Neural Biology, School of Life Sciences, Shanghai University
99 Shang Da Road, Shanghai 200444, China
jo717@shu.edu.cn*

Yong Zhang

*Pulmonary Department, Zhongshan Hospital, Fudan University
180 Fenglin Road, Shanghai 200032, China
zhang.yong@zs-hospital.sh.cn*

Yanhui Li

*Laboratory of Molecular Neural Biology, School of Life Sciences, Shanghai University
99 Shang Da Road, Shanghai 200444, China
liyanhui@shu.edu.cn*

Qing Nie

*Department of Mathematics, University of California
Irvine, Irvine, California, United States of America
qnie@math.uci.edu*

Many major diseases, including various types of cancer, are increasingly threatening human health. However, the mechanisms of the dynamic processes underlying these diseases remain ambiguous. From the holistic perspective of systems science, complex biological networks can reveal biological phenomena. Changes among networks in different states influence the direction of living organisms. The identification of the kernel differential subgraph (KDS) that leads to drastic changes is critical. The existing studies contribute to the identification of a KDS in networks with the same nodes; however, networks in different states involve the disappearance of some nodes or the appearance of some new nodes. In this paper, we propose a new topology-based KDS (TKDS) method to explore the core module from gene regulatory networks with different nodes in this process. For the common nodes, the TKDS method considers the differential value (D-value) of the topological change. For the different nodes, TKDS identifies the most similar gene pairs and computes the D-value. Hence, TKDS discovers the essential KDS, which considers the relationships between the same nodes as well as different nodes. After applying this method to non-small cell lung cancer (NSCLC), we identified 30 genes that are most likely related to NSCLC and extracted the KDSs in both the cancer and normal states. Two significance functional modules were revealed, and gene ontology (GO) analyses and literature mining indicated that

the KDSs are essential to the processes in NSCLC. In addition, compared with existing methods, TKDS provides a unique perspective in identifying particular genes and KDSs related to NSCLC. Moreover, TKDS has the potential to predict other critical disease-related genes and modules.

Key words: kernel differential subgraph, complex biological networks, dynamic process, topology differential value

1 Introduction

Over the past few decades, cancer has become one of the leading diseases causing human death. According to World Cancer Research Fund International (WCRFI)[1], in 2012 alone, the number of new cancer cases was 14.1 million, and this number is expected to reach 24 million by 2035. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer and accounts for 80%~85% of lung cancers. However, there is no effective treatment. Recently, medical professionals and scientific researchers began to explore the dynamic process underlying cancer, but a clear understanding of the mechanisms is lacking. Genes perform their functions and rely on frequent gene-gene interactions throughout the entire system. Networks provide a holistic perspective of a gene's interactome. Generally, a network is a common data structure that can be applied to most biological systems and represents the overall physical and functional components of the system.

Being an important biological network, gene regulatory networks (GRNs) reflect gene changes due to variations in time, space, or the external environment and represent the dynamic regulation among genes. Bagging C3NET (BC3NET) is an excellent algorithm for constructing gene regulatory networks (GRN) from large-scale gene expression data. From several copies of independent bootstrap samples, C3NET, which corresponds to a Bayesian approach with non-informative priors, deduces N control gene networks. Then, BC3NET aggregates the N networks and outputs a weighted network. This method performs well in gene regulatory network inference[2]. Genes in the network become up-regulated or down-regulate due to cancer, ageing and death of the life body. Therefore, it is necessary to extract the kernel differential subgraph (KDS), which is the central module leading to network differential and system mutations, to compare the networks in different states. Network comparisons usually include similarity alignment and differential alignment.

Similarity alignment aims to identify the most similar modules in two different networks. We can identify conserved functional modules across two or more networks at different stages. We can also determine the most similar modules across different species, such as yeast and human. By combining network topology with node (i.e., proteins and genes) sequences across species, an adaptive hybrid algorithm that combines the Hungarian algorithm and the greedy algorithm for the global alignment (HGA) has yielded exciting results[3]. Graphlet[4] based on network topologies and neighbourhoods define a series of non-isomorphic subgraphs and graphlet vectors, which are used to identify similar clusters in networks[5, 6]. An extended graphlet degree distribution agreement using 14 eukaryotic protein-protein interaction (PPI) networks was shown to produce a better result[7].

Differential network comparisons aim to identify the most differential modules causing diversity by integrating dynamic gene expression changes and metabolic fluxes. To achieve this goal, a growing number of studies have focused on the dynamic changes between a pair of genes due to changing conditions, time or species variations. These emergent studies are more concerned with which system modules have changed rather than which modules have remained stable[8]. Bai Zhang proposed the differential dependency network (DNN) method, which is a local dependency model-based method, to detect topological changes across different biological conditions[9]. The knowledge-based differential dependency network (KDDN) method is an extended version that combines biological knowledge with network construction and comparison[10]. DINGO[11] is based on the pathway method and decomposes networks into group-specific and global components to estimate the group-specific conditional dependency relationships. These methods have achieved great success; however, these methods only analyse different networks with the same nodes and different regulatory relationships while ignoring the interactions among the different nodes. During disease development, molecules (e.g., genes and proteins) consistently play crucial roles through their interactions, and the expression of these genes is always different (up-regulated or down-regulated); thus, it is very important to dig these dynamic network biomarkers[12] through differential network comparison.

In this study, we propose a new topology-based KDS (TKDS) method to infer the KDS from gene regulatory networks with different nodes from omics datasets. The TKDS method considers the topology differential value (D-value) between the same gene pair and across different gene pairs. The networks are constructed separately in both the cancer and normal states to analyse the gene expression patterns. Using a baseline of topologically maximized D-values, the TKDS revealed 30 genes that were most likely to be related to NSCLC. The KDS analysis was performed in different states, and a gene ontology (GO)

analyses was also performed to provide a better interpretation of the biological significance of the genes and the dynamic process.

This paper is organized as follows. Section 2 briefly describe the relate work. Section 3 introduces the TKDS method, which can be used to interpret the dynamic process of cancer. The results of the experiment including the constructed network, top differential nodes and KDSs, and the comparative experiment are presented in section 4. We summarize the results of the GO analysis in section 5, and finally, we conclude our experiments and provide possible future directions.

2 Related work

2.1 HGA

HGA is an adaptive hybrid algorithm that combines the Hungarian algorithm and the greedy algorithm for the global alignment of biomolecular networks [3]. The general idea is to iterate over the topological and the biological similarities of the nodes and use the Hungarian algorithm and the greedy algorithm to identify the best mapping between the networks through an iterative process. Fig. 1 illustrates the main process of HGA. The sequence similarity is computed by BLAST[13], which is a common tool for comparing nucleotide and amino acid sequences. The initial similarity matrix is computed by the sequence similarity only. Considering the topologies of each pair of nodes, HGA updates the similarity matrix. For each similarity matrix, HGA performs mapping between the two networks and iterates this process until the results converge by computing the edge correctness (EC)[14] and the points and edges (PE)[15] score.

Fig. 1. Main process of HGA

EC scores are typically used to measure degree of topological similarity and can be estimated as a percentage of matched edges. The EC formula is defined as follows:

$$EC = \frac{\left| \left\{ (u_i, u_j) \in E_A \cap (\Phi(u_i), \Phi(u_j)) \in E_B \right\} \right|}{E_A} \times 100\% \quad (1)$$

where (u_i, u_j) represents the edge between node u_i and u_j in network A; $\Phi(u_i)$ and $\Phi(u_j)$ mean the mapping nodes of u_i and u_j , respectively. $(\Phi(u_i), \Phi(u_j))$ represents the mapping edge in network B; and E_A and E_B represent the edge set of networks A and B, respectively.

PE is defined similarly to the measures used by INM[15], which considers both nodes and edges in the score. In addition, $wt(e_{ij}^A)$ is the weight of edge e_{ij}^A between nodes u_i and u_j in network A, which indicates the reliability of this edge. $sim(u_i, \Phi(u_i))$ is the initial similarity score of u_i and $\Phi(u_i)$. PE is thus defined as follows. The score for an edge (the edge score, ES) equals zero if any of its nodes does not match with its similar nodes and that the score for a node (the point score, PS) equals zero if none of its edges has a score. Therefore, PE is more stringent than EC by considering both nodes and edges.

$$PE = ES + PS = \frac{1}{2} \sum_{i,j=1}^n ES(e_{ij}^A) + \sum_{i=1}^n PS(u_i) \quad (2)$$

where

$$ES(e_{ij}^A) = \begin{cases} wt(e_{ij}^A), & \exists e_{\Phi(u_i), \Phi(u_j)} \in E_B, \\ & \text{and } sim(u_i, \Phi(u_i)) > 0, \\ & \text{and } sim(u_j, \Phi(u_j)) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

$$PS = \begin{cases} sim(u_i, \Phi(u_i)), & \sum_{j=1}^n ES(e_{ij}^A) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

2.2 Graphlet and graphlet vector

Graphlet[4, 5, 7] is a set of non-isomorphic subnets containing 2, 3, 4 or more nodes. As listed in Fig. 2, the graphlets of 2~4 nodes are easy to understand. For each graphlet, nodes with the same topology (degree) are indicated in the same colour and vice versa. Therefore, we can obtain 15 different orbits labelled 0~14. For a specific node in a network, we consider all combinations and obtain the frequency in 15 dimensions; we refer to these combinations as graphlet vectors.

As shown in Fig. 2, noticing the redundancy in dimensional information between these graphlets (14 with 3, 7 with 2), Tijana et al. introduced a weight for each dimension of the graphlet vector. Higher dimensions typically contain the smaller dimensions.

The graphlet vector elaborately depicts the topological relation between a node and its specified number of nearest neighbours. When the local structure changes in a network, the graphlet vector for relative nodes changes accordingly. The more the local structure changes, the more likely it becomes that the primary graphlet vector varies as well.

Fig. 2. 2~4-node graphlets $G_0, G_1, \dots, G_8$ and their automorphism orbits 0, 1, ..., 14[4].

3 Topology-based Kernel Differential Subgraph (TKDS)

3.1 Problem formulation

Biomolecular networks are a type of data structure in graph theory. The nodes in the networks represent proteins, genes or other molecules, and the edges denote the connections of each pair of node. For example, in GRNs, the nodes are the genes, and the directed or undirected edges are the interactions among these genes. The purpose of assessing life processes through network comparison is to identify the maximized KDS. The alignment process is described as follows.

Networks in two states are represented as $G(V, E)$ and $G'(V', E')$, where V and V' represent the nodes set; E and E' represent the edges set; and $N(V)$, $N(V')$, $N(E)$, and $N(E')$ represent the numbers of nodes and edges, respectively, in networks G and G' . Generally, the number of nodes in the two networks may be different. The common nodes between the two networks are represented as $CommN(G, G')$. To compare the topological changes between the two networks, the same nodes are grouped together. For different nodes in the smaller network, the best matching nodes in the other relatively larger network are identified.

For the two given networks, the D-value of node $u \in V, u' \in V'$ is defined as $D_{uu'}$. However, the nodes of the two networks must be identical. Actually, when we apply dynamic network comparison to the networks, we cannot guarantee the abovementioned condition. Using HGA and node sequence similarity $s(u, u')$, we easily overcome this limitation. For every pair of nodes, we redefine the topology differential degree value as follows:

$$D_{uu'} = \frac{\sum_{k=0}^{14} \omega_k (\log(\text{vec}_k(u) + 1) - \log(\text{vec}_k(u') + 1))}{\sum_{k=0}^{14} \omega_k \log(\max\{\text{vec}_k(u), \text{vec}_k(u')\} + 2)} * s(u, u') \quad (5)$$

where $\text{vec}_k(u)$ and $\text{vec}_k(u')$ represent the graphlet vector in the k -dimension for the node pair (u, u') in the two networks $G(V, E)$ and $G(V', E')$ and ω_k is the weight of each dimension in the graphlet vector. Notably, when the node pair (u, u') mapped by HGA is $(u = u')$, the similarity $s(u, u') = 1$. Then, across the top D-value nodes, the KDS can be extracted using the proposed TKDS method.

In this paper, the TKDS method consists of several steps, including gene selection, network construction, alignment, analysis and further studies. The framework of the TKDS algorithm is depicted in Fig. 3.

Fig. 3. Framework of the TKDS algorithm

3.2 Gene selection and network construction

The original data provided in the NCBI GEO database are unprocessed. One gene chip usually contains millions of probes. Probes are obtained from different companies, and the distribution of the reference sequences varies. To construct a network in distinct states, we must pre-process the data by performing the following three steps: background correction, control-based normalization and summarization. In this study, we used the RMA[16] standardized method in Bioconductor to construct the gene expression matrix. Originally, Bioconductor was exclusively used for analysing gene chips. Considering the effectiveness of using prior knowledge[17], we referred to the disease-related pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[18] for gene selection. Then, a Pearson correlation-based method was used to screen the genes that have strong relationships with the genes in the mentioned pathway. By controlling the correlation coefficient, we obtained a certain number of genes in the disease or normal state. Then, we extracted the genes' expression matrix as X and X' . We used BC3NET to construct the two networks $G(V, E)$ and $G'(V', E')$ based on the gene expression matrix.

3.3 Network pairwise alignment

HGA is used to align the smaller network $G(V, E)$ to the larger network $G'(V', E')$. For $u \in G, u', v' \in G'$, where u and u' indicate the same gene in G and G' , respectively, v' is the most similar gene in G' . Two matching relationships are noted as follows:

$$u \rightarrow u', \text{ if } u \in G' \quad (6)$$

$$u \rightarrow v', \text{ if } u \notin G' \quad (7)$$

Thus, we obtained network $G^*(V^*, E^*)$, where $V^* \subseteq V'$, $E^* \subseteq E'$ and $N(V^*) = N(V)$.

3.4 Differential gene extraction

As described in section 3.3, when we match the two GRNs, each node in the small network G matches to a corresponding node in G' , and the mismatched nodes in G' are deleted. In fact, there are fewer nodes without matches because we can control the scale of the larger network G' before using

HGA for the alignment. For example, we can control the scale by choosing a subnet that includes the genes in a pathway and their first, second or third neighbours. For the same node pairs (u, u') , we can calculate the D-value $D(u, u')$ according to its topology distribution. For different node pairs (u, v') that belong to the relationship in (7), we also consider the nucleotide sequence similarity $s(u, v')$. Tijana et al. proposed that u and v' change appreciably if $D(u, v') > 0.4$ [5]. Thus, we selected the appreciably changing gene set $Diff$, and these genes represent the most important genes across the dynamic process.

3.5 Kernel differential subgraph extracting

We obtained $Diff$ from the two original networks. To analyse the relationships between these nodes, we need to identify their minimal connected subgraph, which is similar to the Steiner tree problem. We consider the connectivity between the given nodes in $Diff$ and the other nodes in network G/G^* , and TKDS uses the sorted D-value genes to extract the KDS. Algorithm 1 describes the extraction procedure:

Algorithm 1: Get KDS

Input: Network G , differential nodes $Diff$ with their D-value

Output: KDS of G

Sort $Diff$ by their D-value

$G_x \leftarrow G$

foreach $node1$ **in** $Diff$ **do**

$neighbours \leftarrow getNeighbours(G_x, node1)$

 Sort $neighbours$ by their D-value

foreach $node2$ **in** $neighbours$ **do**

 deleteEdge($G_x, node1, node2$)

if notConnected($Diff$) **then**

 addEdge($G_x, node1, node2$)

end

end

deleteIsolatedNodes(G_x)

$KDS \leftarrow Subnet(G, Nodes(G_x))$ /* get subnet KDS of left nodes in G_x from G */

Here, we illustrate the process of KDS abstraction as a toy example. Suppose there are 5 nodes in the original network G and G^* as shown in Fig. 4. The hypothetical D-values of these 5 pairs of nodes are listed in Table 1.

Fig. 4. Toy example of the networks G and G^* . The red nodes represent the nodes in $Diff$, whereas the blue nodes are not in $Diff$.

Table 1 Hypothetical D-values of the 5 pairs of nodes between G and G^*

Nodes in G	A	D	E	B	C
Nodes in G^*	A'	D'	E'	H'	F'
D-value	0.70	0.50	0.45	0.35	0.28

According to the steps outlined in algorithm 1, the removal order of these edges in G is $(E, B) \rightarrow (A, B)$ and in G^* it is $(E', F') \rightarrow (E', H') \rightarrow (D', F')$. Therefore, the final results of the KDS in G and G^* are as shown in Fig. 5.

Fig. 5. KDS of G and G^* of the toy example.

Finally, the KDS of G/G^* was obtained, which includes the differential nodes $Diff$. To further determine the dissimilarity between the two KDSs, we compute the PE and EC scores. In this study, we also evaluated the p-values of PE and EC in a KDS with 100-time random results. The random connected

subnet was generated using a specified number of genes. The validation of the dissimilarity between the results produced by the TKDS algorithm and random results is a frequently used statistical method. This dissimilarity is characterized by the p -value. The smaller the p -value is, the more reliable the result is.

4 Experiments

4.1 Dataset and a priori knowledge

A classic dataset was used in this work to analyse the molecular signature of non-smoking females with NSCLC in Taiwan[19]. This dataset contains gene expression analyses of RNA extracted from paired tumour and normal tissues using the Affymetrix Human Genome U133 Plus 2.0 Array. The dataset includes 60 cancerous samples and 60 normal samples. Only one pathway, i.e., hsa05223 in NSCLC of Homo sapiens, is found in the KEGG database, and 58 genes are identified. The gene nucleotide sequences are retrieved from Ensembl using EMBL-EBI[20], which freely provides abundant biological data and online services.

4.2 Network construction and pairwise alignment

First, we selected 58 genes in the NSCLC-related pathway and other genes with strong Pearson correlations (e.g., 0.8 or greater) according to the pre-processed gene expression matrix. Then, 186 and 192 genes were retained in the cancer and normal states, respectively. Subsequently, using the BC3NET algorithm (in which the bootstrap is set to 100, the estimator is set to Pearson and the significance level is set to 1.05), the original cancer and normal networks $G(V, E)$ and $G'(V', E')$ were constructed, where $V=186$, $E=529$, $V'=192$, and $E'=524$. Fortunately, there are 70 genes in both the cancer and normal network. Thus, 58 genes in the pathway are involved; furthermore, 12 correlated genes are also involved. For the 70 common genes, the HGA matched these genes with themselves, and the other 116 genes in G were matched to the most similar genes in G' . Thus, the $G'(V', E')$ becomes the subnet $G^*(V^*, E^*)$, where $V^*=V=186$ and $E^*=504$.

4.3 Topological differential value

The D-values of the common genes are the foundation of the analysis of the dynamic changes in the networks. Table 2 displays the D-values of these genes in the two networks. Notably, the genes PRKCG, NRAS, HRAS, RXRA, and RASSF1 exhibit great significance according to graphlet theory. These genes have much higher D-values than the other genes and are the focus of our study. According to Tijana et al., 30 genes with differential values greater than 0.4[5] are selected for further analysis. A functional annotation of these genes using DAVID[21] revealed that 18 genes cause diseases, and RASSF1, CCND1, CDK4, CDK6, EFG, EFGR, FHIT and HRAS are highly related to cancer. From recent studies[22], 22 of the 30 genes (HRAS, SOS1, CDK6, RXRA, EGFR, RXRG, PIK3R3, ARAF, EGF, RASSF1, PLCG1, RXRB, CDK4, CCND1, NRAS, PIK3CD, PRKCG, STK4, FHIT, E2F2, PRKCA and AKT3) are closely related to the incidence of NSCLC.

Table 2. D-values of the common genes

#	Gene name	D-value	#	Gene name	D-value	#	Gene name	D-value
1	PRKCG	0.778963	25	CDK4	0.442507	48	PIK3R1	0.290994
2	NRAS	0.695101	26	CCND1	0.433237	49	CD46	0.272574
3	HRAS	0.635034	27	PRKCA	0.424355	50	E2F3	0.270518
4	RXRA	0.630949	28	PIK3R3	0.405049	51	SOS2	0.267528
5	RASSF1	0.612386	29	AKT3	0.403718	52	RB1	0.26464
6	RXRB	0.612292	30	EGFR	0.401243	53	BRAF	0.254285
7	ALK	0.600895	31	STAT3	0.383504	54	RARB	0.250133
8	GNAI3	0.591487	32	BAD	0.373768	55	KRAS	0.242789
9	EGF	0.589939	33	AKT1	0.364487	56	JAK3	0.242788
10	E2F2	0.589505	34	E2F1	0.363187	57	HIPK3	0.237009
11	SLC6A19	0.587554	35	PRKCB	0.363169	58	PTPN11	0.215192
12	TMEM259	0.571678	36	RAF1	0.343301	59	CHM	0.213873
13	UTRN	0.529541	37	MAP2K2	0.342685	60	PLCG2	0.211704
14	STAT5A	0.516495	38	PDPK1	0.340763	61	MAP2K1	0.206533
15	ARAF	0.496771	39	MAPK1	0.335902	62	CDKN2A	0.206364
16	PLCG1	0.475403	40	MAPK3	0.324001	63	TP53	0.174156
17	SOS1	0.473332	41	AKT2	0.317212	64	PIK3CA	0.171094
18	CLDND1	0.471871	42	PIK3CB	0.314745	65	GTF2I	0.163353
19	FHIT	0.471599	43	STAT5B	0.313158	66	PSEN1	0.154414
20	CDK6	0.470846	44	RASSF5	0.310598	67	ERBB2	0.148002
21	RXRG	0.454058	45	TGFA	0.299917	68	PIK3R2	0.142001
22	STK4	0.452936	46	MFS14A	0.296712	69	FOXO3	0.1093
23	EML4	0.452345	47	CASP9	0.294937	70	GLYR1	0.108372
24	PIK3CD	0.449509						

Spinal cord cerebellar ataxia type 14 is associated with a PRKCG mutation. Gene mutations in PRKCG promote the occurrence of spinal cord cerebellar ataxia type 14. Additionally, the PRKCG gene plays crucial roles in tumour promotion, cell proliferation, differentiation, and migration by encoding γ PKC[23, 24]. As shown in Table 2, many changes occurred in the topology of PRKCG.

The RAS gene is a proto-oncogene, and a mutation in this gene decreases its endogenous free guanylic acid phosphatase (GTPase) activity and reduces the binding of the GTPase activating protein [25]. Consequently, the mutation strengthens the combined effect of the RAS proteins and CTP and contributes to cell proliferation and malignancy transformation. The epidermal growth factor receptor (EGFR) type I transmembrane receptor tyrosine kinase is a proto-oncogene C-erbB1 expression product, and NRAS and HRAS are two members of the RAS gene family[26].

EGFR was overexpressed in a variety of epithelial malignant tumours, such as NSCLC, colon cancer, head and neck cancer and ovarian cancer[27]. Moreover, the expression level of EGFR was closely related to tumour cell proliferation, adhesion, invasion and angiogenesis. RASSF1A was discovered as a new tumour suppressor gene, and methylation of its promoter region influences the occurrence and development of gastrointestinal tumours. By blocking the accumulation of Cyclin D1 and controlling the G1/S phase, RASSF1A can inhibit tumour cell growth[28]. E2F2 is a member of the E2F family and is a nuclear transcription factor; E2F2 can regulate cell mitosis and control cell growth, differentiation, and proliferation by regulating the cell division cycle-related genes[29].

4.4 Differential kernel subgraph results

The topology of the abovementioned genes has clearly changed, which explains the importance of physical properties. In biology, genes execute their functions within interacting functional modules. Thus, the TKDS extracts the KDSs $S(G)$ and $S(G^*)$ of the 30 differential genes from the original networks G and G^* . The results are visualized using Gephi[30] as shown in Fig. 6: the blue nodes represent 30 genes that are common to both the cancer and normal states, the red nodes represent the differential genes connected to the 30 genes represented in blue, the green nodes represent the common nodes added using the proposed method, and the small black nodes are the ordinary genes that do not exist in the KDS. A and B represent the overall network, and C and D are the two extracted KDSs in the cancer and normal states, providing a more distinct visualization. Every node and its counterpart are at the same location in these four networks

In addition to the 70 common genes, TKDS considers the relationships between these genes and the genes that are exclusively present in the cancer network and constructs the kernel differential subgraph (KDS) in cancer state with 21 additional genes added. Interestingly, by searching the COSMIC database[31], we found 3 of the 21 genes that cause non-small cell lung cancer (NSCLC). In addition, two subnets are obtained from the 3 genes and their neighbours. As noted in Fig. 7, there are two subnets in the cancer state containing the newly added genes PIK3CB, MAP2K2 and AKT1.

Fig. 6. Overall network and isolated KDS in the cancer and normal states. The blue nodes represent 30 genes that are common to both the cancer and normal states, the red nodes represent the differential genes that connect to the 30 genes represented in blue; the green nodes represent the common nodes added using the proposed method, and the small black nodes are the ordinary genes that do not exist in the KDS.

The above two functional modules mainly relate to important molecules of the PI3K/AKT, MAPK and PRKC signalling pathways. The PI3K/AKT signal transduction pathway plays an important role in the regulation of cell survival and proliferation. PI3K/AKT signalling dysfunction plays a crucial role in multiple tumourigenesis and development models and is associated with tumour anti-apoptosis, migration and invasion[32].

Fig. 7. Functional modules in the KDS containing newly added genes related to NSCLC

Studies have demonstrated abnormal up-regulation of PI3K in breast cancer, prostate cancer, ovarian cancer, colorectal cancer and other human malignancies[33]. Class 1A phosphatidylinositol kinases (PI3Ks) are important core molecules in the PI3K/AKT signal transduction pathway that are mainly activated by various kinases and G protein-coupled receptors. PI3K is a lipid kinase family consisting of a catalytic subunit (PIK3C) and a regulatory subunit (PIK3R), in which the catalytic subunit contains four members. PIK3CB protein is widely abundant in the cell, regulates synthesis and promotes the normal development of cells. The deletion of the PIK3CB gene leads to the death of embryonic cells. Knockdown of the catalytic subunit PIK3CB reduces the expression of proliferating regulatory protein E and Bcl-2, and the expression levels of genes downstream of PI3K, including AKT1, AKT2, pAKT and mTOR, are also significantly decreased. PIK3CB mutations result in the overexpression of its encoded protein PI3Kp110 β , enhanced PI3K catalytic activity, and activation of the PI3K/AKT signalling pathway[34], leading to cell metabolism disorders and cell carcinogenesis[35].

The PI3K/AKT pathway integrates numerous signal pathways to form complex signal networks. Mitogen-activated protein kinases (MAPKs) are an intracellular class of serine/threonine protein kinases. Studies have demonstrated that the MAPK signal transduction pathway exists in the majority of cells and plays a crucial role in the signal transduction of extracellular stimulation signals into cells and nuclei and cellular biological responses (such as cell proliferation, differentiation, transformation and apoptosis,). The SOS1 gene was discovered in 2002 and represents one of the successfully cloned and identified HGF pathogenicity genes[36]. SOS1 imports exogenous genes into the nucleus to regulate cell proliferation

through interaction with the MAPK pathway. SOS1 gene knockout or inhibition of SOS1 gene expression prevents the proliferation of tumour cells and even leads to tumour cell apoptosis. Growth factors play an important regulatory role in cell division in multicellular development processes. The phospholipase C- γ 1 (PLC- γ 1) and phosphatidylinositol-3 kinase (PI- 3K) are two important signal mediators in this pathway, which can be activated by growth factor receptors and then trigger the next step of signalling transmission. Genes involved in the module participate in cell proliferation, differentiation, migration and invasion, which are associated with tumourigenesis and development.

To further determine whether the KDS is more differential than the original network, we compute the PE and EC scores for the original network G and G^* and the KDS $S(G)$ and $S(G^*)$. Table 3 lists the values obtained using the different methods. PE and EC show, to a certain extent, the similarity of the two networks. The TKDS PE score of 7.86 and EC score of 0.02 are lower than the original network PE score of 60.08 and EC score of 0.16. Thus, the KDS is more differential than the original network. Thus, the more differential, the more important it is to reveal the dynamic change between different states. Next, the average PE and EC scores of 100 randomly connected subnets are calculated. The number of genes in the random subnets is equal to that in the differential subnet obtained using the proposed method. The TKDS PE score of 7.86 and the EC score of 0.02 are lower than the scores obtained using the random method. Therefore, the proposed TKDS method is superior to the random method.

Table 3 PE and EC scores obtained using different methods

Item	Nodes in cancer	Nodes in normal state	PE score	EC score
Original	186	192	60.08	0.16
TKDS	51	44	7.86	0.02
Random	51	44	12.65	0.14
Shortest path-based method	67	54	7.51	0.10

Certainly, we hope to identify the lowest KDSs possible and enhance the difference as much as possible. Consequently, we propose an intuitive shortest path-based method to construct the network and union of these paths in a new subgraph. Because of the non-uniqueness of the shortest path, we repeat this path 50 times. As shown in Table 3, the TKDS PE score of 7.86 is similar to the shortest path-based method score of 7.51, and the EC score of 0.02 is lower than the score of 0.10 obtained using the shortest path-based method. Furthermore, the average number of nodes is 67 and 54 using the shortest path-based method, which is greater than the number obtained using the TKDS (51 and 44).

Fig. 8. PE and EC score distributions obtained using different methods. A and B represent the distributions of the PE and EC scores, respectively, obtained with the random method. C and D display the tendencies of the PE and EC scores, respectively, obtained with the shortest-path-based method.

Fig. 8 A and B present the distribution of the PE and EC values in the random subnet, and the p -values calculated using single sample t -tests in SPSS are less than 0.01. The distribution of the PE and

EC values using the proposed and the shortest path-based methods are shown in Fig. 8 C and D, which further illustrate the effectiveness of the proposed method.

4.5 GO analyses

GO analyses have been widely used to functionally define and describe genes and proteins. In this study, we mapped these differentially changing genes with D-values greater than 0.4 to the GO items of biological process (BP), molecular function (MF) and cellular component (CC) using the Gene Ontology Consortium database[37, 38]. Fig. 9 presents the categories of the GO functions of the genes with the top 30 D-values. As shown below, the four leading processes of response to stimulations, biological regulation, metabolic process and cell communication are the most enriched with the formation and development of a tumour. Membrane and nuclear genes account for a large proportion of the cellular component, and protein binding plays a key role in functional molecular processes[39].

Fig. 9. GO enrichment items in the 30 differential genes but not in the 70 common genes. The black bar represents the number of differential genes, and the red, blue and orange bars represent the BP, MF and CC enrichment results of GO analysis, respectively.

Fig. 10. GO enrichment items in the 30 differential genes but not the 70 common genes. The black rectangle represents the number of the differential genes. The red, blue and orange rectangles represent the BP, MF and CC enrichment results of the GO analysis, respectively.

Different GO items were then obtained by comparing the 30 differential genes with the original 70 common genes. As shown in Fig. 10, the top enriched items of the three GO categories are listed; the descriptions of the different GO items are provided in the appendix. The enrichment was mainly focused on the regulation of cell cycle, activity and function of protein kinases, as well as the replication of nucleic acid replication, which are critical for the development of cancer. The regulation of cell cycle is a complex biological process. It has been reported that the dysregulation of the cell cycle is closely related to the transformation of normal cells into tumour cells. Tumours represent a type of disease involving the disruption of regulation mechanisms underlying the cell cycle, as there are large changes in genes that regulate the cell cycle in tumour cells. The common protein kinases included mitogen-activated

protein kinase (MAPK), protein kinase C (PKC), PKA, and calcium/calmodulin-dependent protein kinase (CAMK). The mitogen-activated protein kinase kinase 4 (MKK4) is the key convergence point and critical regulator in the MAPK signal pathway. MKK4 can directly or indirectly regulate a variety of crucial signalling transduction molecules inside and outside the nucleus, participating in apoptosis, inflammation, cell proliferation and other processes and ultimately promoting the occurrence of tumours. Tyrosine protein kinase plays an important role in the intracellular signal transduction pathway, and regulates a series of physiological processes, including cell growth, differentiation and death; moreover, it is also associated with the proliferation, differentiation, migration and apoptosis of tumour cells. Activated protein kinase C (PKC) is widely involved in cellular information transfer, secretion, regulation of ion channels, cell proliferation, differentiation, and a series of life processes.

Fig. 11. GO enrichment in the cancer KDS but not in the normal state KDS. The black bar represents the number of genes in the KDS of cancer, and the red, blue and orange bars represent the BP, MF and CC enrichment results of the GO analysis, respectively.

If several genes in a cluster share one or more GO terms, the gene cluster is believed to be biologically significant. The GO enrichment analysis of BP, MF and CC in the genes in the cancer and normal subnets (G and G*, respectively) are shown in Fig. 11. These GO terms are enriched in the cancer state not in the normal state. In the cancer state, 32 genes are associated with ion binding in the analysis of the different GO categories. An ion channel is a hydrophilic microporous protein channel that makes the cell membrane permeable. Ion channels are important components for maintaining cell activity, and they are also closely related to tumour development. The abnormal expression and change in activity in tumour cell ion channels regulate tumour cell proliferation and apoptosis. Potassium and calcium ion channels are involved in tumour angiogenesis, and sodium and chloride ion channels regulate the invasion and metastasis of tumours. The voltage-gated sodium channel could influence the intracellular Ca^{2+} concentration, voltage gradient and conversion to the transient polarization state, which is associated with malignancy and metastasis of tumour cells[40].

Cell metabolism, macromolecular metabolism and protein metabolism are all enriched, which demonstrates that tumour metabolism is important for tumourigenesis and development. The activation of the PI3K-AKT pathway promotes the expression of the glucose transporter GLUT1 on the cell membrane surface, contributing to the glycolysis of tumour cells and tumour malignancy[41, 42]. The hypoxic tumour microenvironment, oncogene activation, tumour suppressor gene inactivation, oxidative phosphorylation damage, inflammation, and microRNAs all contribute to tumour cells and the glycolysis metabolic phenotype, which is used to help tumour cells adapt to environmental pressures[43, 44]. A series of phenotypic tumour formation changes, such as the lack of cell movement and contact-mediated inhibition of cell proliferation, reduction in cell surface viscosity, appearance of cancer specific antigens, and invading, resisting or avoiding phenomena, such as the host's immune responses, were identified. Altogether, these processes are closely related to membrane structure and function.

4.4 Performance comparison with kDDN

The differences between networks of different conditions, such as stages of disease development, occur as a result of changes in the common nodes (genes, proteins, etc.) in these networks and also depend on the disappearance or the appearance of some nodes. At present, studies on network differences mainly focus on the networks with the exact same nodes, not networks with some different nodes. The TKDS

method considers topological differences not only between the common nodes, which are described by D-values, but also different nodes that are omitted by other methods. In addition, TKDS considers the relationships between common and different nodes. As previously described, kDDN is an excellent algorithm for comparing differences in networks with the same nodes. We therefore compared our topology-based method TKDS with kDDN.

In the 70 common genes, after computing the centrality of the network response to cancerization, 13 overlapping genes in the top 30 differential genes were identified by TKDS (top30-TKDS) and kDDN (top30-kDDN). Both methods identified the following 13 genes: SOS1, CDK4, ALK, STK4, ARAF, NRAS, HRAS, CLDND1, EGFR, AKT3, E2F2, PRKCG and PIK3CD. Moreover, the top30-TKDS are all related to cancer as demonstrated by literature mining. The GO items that are exclusively enriched by the top30-TKDS are listed as follows in Fig. 12.

Fig. 12. GO items that are exclusively enriched by the top30-TKDS. The red, blue and orange bars represent the number of genes enriched in the BP, MF and CC analyses, respectively.

Positive regulation of cellular component movement (GO:0051272), migration (GO:0030335), motility (GO:2000147), locomotion (GO:0040017) and proliferation (GO:0050679) are highly important in cancer cell cleavage and proliferation[39, 45]. Regarding the 70 common genes, TKDS considers the relationships between these genes and the genes that are exclusively in the cancer network and constructs the kernel differential subgraph (KDS) in the cancer state with 21 additional genes added. Interestingly, by searching the COSMIC database, we identified 3 of the 21 genes that cause non-small cell lung cancer (NSCLC). Two subnets are obtained from the 3 genes and their neighbours. The two functional modules shown in Fig. 7 are mainly related to important molecules of the PI3K/AKT, MAPK and PRKC signalling pathways which play critical roles in the follow-up studies of NSCLC.

5 Conclusions

In this study, we proposed the TKDS method, a topology-based method, to infer the KDS from gene regulatory networks with different nodes. For the common nodes, the TKDS method considers the D-value of the topological change. For the different nodes, TKDS identifies the most similar gene pairs and computes the D-value. Hence, TKDS discovers the essential KDS, which considers the relationships between the same nodes as well as different nodes.

After applying the TKDS in non-small cell lung cancer gene array dataset. We extracted the KDS containing the abovementioned genes with larger D-values using the TKDS algorithm and two functional modules related to NSCLC were also discovered. Further analyses, including calculation of the PE and EC scores, reviewing the literature and performing a GO analysis, confirmed the reliability of the results. Compared with the published method kDDN, TKDS identified both the apparent genes related to NSCLC and two essential functional modules in the process. Further GO analysis also demonstrated the effectiveness of our proposed method. Studies investigating NSCLC are continuously performed, and the results of this paper open the door to a new world in differential network alignment. These findings may inspire solutions to NSCLC research-related limitations and provide new target genes that were previously undiscovered.

6 Acknowledgements

This work was partially supported by the National Key R&D Program of China [No. 2017YFB0701501], the Project of NSFS [No. 17ZR1409900], the National Natural Science Foundation of China [No. 31500827], and the Project of Young Eastern Scholar [NO. QD2015033].

7 References

1. Ferlay, J., et al., *GLOBOCAN 2012. Cancer incidence and mortality worldwide: IARC Cancer Base No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer 2013.[Visitado 2014 Dic 22]*. 2016.
2. Emmert-Streib, F., et al., *The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks*. *Frontiers in genetics*, 2014. **5**.
3. Xie, J., et al., *An adaptive hybrid algorithm for global network alignment*. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016. **13**(3): p. 483-493.
4. Pržulj, N., D.G. Corneil, and I. Jurisica, *Modeling interactome: scale-free or geometric?* *Bioinformatics*, 2004. **20**(18): p. 3508-3515.
5. Milenkovic, T. and N. Pržulj, *Uncovering biological network function via graphlet degree signatures*. *Cancer Informatics*, 2008. **6**(1): p. 257-273.
6. Malod-Dognin, N., K. Ban, and N. Pržulj, *Unified Alignment of Protein-Protein Interaction Networks*. *Scientific Reports*, 2017. **7**(1): p. 953.
7. Pržulj, N., *Biological network comparison using graphlet degree distribution*. *Bioinformatics*, 2007. **23**(2): p. e177-e183.
8. Ideker, T. and N.J. Krogan, *Differential network biology*. *Molecular systems biology*, 2012. **8**(1): p. 565.
9. Zhang, B., et al., *Differential dependency network analysis to identify condition-specific topological changes in biological networks*. *Bioinformatics*, 2009. **25**(4): p. 526-532.
10. Tian, Y., et al., *Knowledge-fused differential dependency network models for detecting significant rewiring in biological networks*. *BMC systems biology*, 2014. **8**(1): p. 87.
11. Ha, M.J., V. Baladandayuthapani, and K.-A. Do, *DINGO: differential network analysis in genomics*. *Bioinformatics*, 2015. **31**(21): p. 3413-3420.
12. Chen, L., et al., *Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers*. *Scientific reports*, 2012. **2**(7391): p. 342.
13. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. *Nucleic acids research*, 1997. **25**(17): p. 3389-3402.
14. Kuchaiev, O. and N. Pržulj, *Integrative network alignment reveals large regions of global network similarity in yeast and human*. *Bioinformatics*, 2011. **27**(10): p. 1390-1396.
15. Xie, J., et al. *A querying method with feedback mechanism for protein interaction network*. 2011. IEEE.
16. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data*. *Nucleic acids research*, 2003. **31**(4): p. e15-e15.
17. Wang, Z., et al., *Incorporating prior knowledge into gene network study*. *Bioinformatics*, 2013. **29**(20): p. 2633-2640.
18. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*. *Nucleic acids research*, 2000. **28**(1): p. 27-30.
19. Lu, T.-P., et al., *Identification of a novel biomarker SEMA5A for non-small cell lung carcinoma*

- in non-smoking women. *Cancer Epidemiology and Prevention Biomarkers*, 2010. **19**(10): p. 2590-2597.
20. Yates, A., et al., *Ensembl 2016*. *Nucleic acids research*, 2015. **44**(Database issue): p. D710-D716.
 21. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. *Nature protocols*, 2009. **4**(1): p. 44-57.
 22. Hosseinzadeh, F., et al., *Prediction of lung tumor types based on protein attributes by machine learning algorithms*. *SpringerPlus*, 2013. **2**(1): p. 2193-1801.
 23. Lu, H., et al., *Genetic variations in the PRKCG gene and osteosarcoma risk in a Chinese population: a case-control study*. *Tumor Biology*, 2015. **36**(7): p. 5241-5247.
 24. Jori, F.P., et al., *Role of RB and RB2/P130 genes in marrow stromal stem cells plasticity*. *Journal of cellular physiology*, 2004. **200**(2): p. 201-212.
 25. Vaucheret, H., C. Béclin, and M. Fagard, *Post-transcriptional gene silencing in plants*. *Journal of cell science*, 2001. **114**(17): p. 3083-3091.
 26. Budhu, A., et al., *Identification of metastasis - related microRNAs in hepatocellular carcinoma*. *Hepatology*, 2008. **47**(3): p. 897-907.
 27. Khan, M.S., et al., *Lack of mutational events of RAS genes in sporadic thyroid cancer but high risk associated with HRAS T81C single nucleotide polymorphism (case-control study)*. *Tumor Biology*, 2013. **34**(1): p. 521-529.
 28. Khanna, C., *Novel targets with potential therapeutic applications in osteosarcoma*. *Current oncology reports*, 2008. **10**(4): p. 350-358.
 29. Sadasivam, S. and J.A. DeCaprio, *The DREAM complex: master coordinator of cell cycle dependent gene expression*. *Nature reviews. Cancer*, 2013. **13**(8): p. 585-595.
 30. Bastian, M., S. Heymann, and M. Jacomy, *Gephi: an open source software for exploring and manipulating networks*. *Icwsn*, 2009. **8**: p. 361-362.
 31. Forbes, S.A., et al., *COSMIC: High - Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer*. *Current protocols in human genetics*, 2016: p. 1-37.
 32. Carnero, A., et al., *The PTEN/PI3K/AKT signalling pathway in cancer, therapeutic implications*. *Current cancer drug targets*, 2008. **8**(3): p. 187-198.
 33. Sirica, A.E., *Role of ErbB family receptor tyrosine kinases in intrahepatic cholangiocarcinoma*. *World journal of gastroenterology: WJG*, 2008. **14**(46): p. 7033-7038.
 34. Yuan, T.L. and L.C. Cantley, *PI3K pathway alterations in cancer: variations on a theme*. *Oncogene*, 2008. **27**(41): p. 5497-5510.
 35. Abubaker, J., et al., *Clinicopathological analysis of colorectal cancers with PIK3CA mutations in Middle Eastern population*. *Oncogene*, 2008. **27**(25): p. 3539-3545.
 36. Hart, T.C., et al., *A mutation in the SOS1 gene causes hereditary gingival fibromatosis type 1*. *The American Journal of Human Genetics*, 2002. **70**(4): p. 943-954.
 37. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology*. *Nature genetics*, 2000. **25**(1): p. 25.
 38. Gene Ontology, C., *Expansion of the Gene Ontology knowledgebase and resources*. *Nucleic acids research*, 2017. **45**(D1): p. D331-D338.
 39. Zhou, Y., et al., *Feature classification and analysis of lung cancer related genes through gene ontology and KEGG pathways*. *Current Bioinformatics*, 2016. **11**(1): p. 40-50.
 40. Fraser, S.P. and L.A. Pardo, *Ion channels: functional expression and therapeutic potential in cancer*. *EMBO reports*, 2008. **9**(6): p. 512-515.

41. Morales, D.R. and A.D. Morris, *Metformin in cancer treatment and prevention*. Annual review of medicine, 2015. **66**: p. 17-29.
42. Lien, E.C., C.A. Lyssiotis, and L.C. Cantley, *Metabolic reprogramming by the PI3K-Akt-mTOR pathway in cancer*, in *Metabolism in Cancer*. 2016, Springer. p. 39-72.
43. Boroughs, L.K. and R.J. DeBerardinis, *Metabolic pathways promoting cancer cell survival and growth*. Nature cell biology, 2015. **17**(4): p. 351-359.
44. Warburg, O., *On the origin of cancer*. Science, 1956. **123**(3191): p. 309-314.
45. Zöschbauer-Müller, S., et al., *Aberrant promoter methylation of multiple genes in non-small cell lung cancers*. Cancer research, 2001. **61**(1): p. 249-255.

Jiang Xie received her PhD in computer science from Shanghai University, China in 2008. She is currently an associate professor in the School of Computer Engineering and Science in Shanghai University. Her main research topics include computational biology, bioinformatics, and high-performance computing.

Dongfang Lu received his BS degree from Yantai University in 2015. He is now studying towards a master's degree in the School of Computer Engineering and Science of Shanghai University. His main research topics are bioinformatics and high-performance computing.

Jiixin Li received his BS degree from China University of Mining and Technology in 2016. He is now studying towards a master's degree in the School of Computer Engineering and Science of Shanghai University. His main research topics are bioinformatics and high-performance computing.

Jiao Wang received her Ph.D. in neural science from Shanghai University, China in 2012. She is currently an assistant professor at the Life Science Centre. Her main research topics include brain research, bioinformatics and neural degeneration disease.

Yong Zhang received his MD. & PH.D in pulmonary department from Fudan University, Shanghai, China in 2013. He is currently associated consultant doctor of pulmonary department in Zhongshan hospital. He is also member and reviewer of Springer, member of Chinese Medical Association. He is experienced in lung cancer screening, also devote to lung cancer biomarker detection and molecular mechanism with genomic sequencing and bioinformatics methods.

Yanhui Li received her BS degree in pharmaceutical engineering from Anhui University of Science and Technology, China in 2015. She is currently a postgraduate at the Life Science Centre. Her main research topics include brain research, bioinformatics and neural degeneration disease.

Qing Nie is a professor in the Department of Mathematics, Department of Biomedical Engineering, and Department of Developmental and Cell Biology at the University of California, Irvine. His key research areas are computational systems biology, cell signalling, stem cell, morphogenesis, and patterning, scientific computing and numerical analysis.