Mathematical Biosciences and Engineering

MBE, 17(4): 3224–3239.
DOI: 10.3934/mbe.2020183
Received: 08 February 2020
Accepted: 16 April 2020
Published: 23 April 2020

http://www.aimspress.com/journal/MBE

Research article

Dynamic gene regulatory network reconstruction and analysis based on clinical transcriptomic data of colorectal cancer

Ancheng Deng[1], Xiaoqiang Sun[2,*]

[1] School of Life Science, Sun Yat-sen University, Guangzhou 510275, China
[2] Key Laboratory of Tropical Disease Control, Chinese Ministry of Education, Zhong-Shan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China

* **Correspondence:** E-mail: sunxq6@mail.sysu.edu.cn.

Abstract: Inferring dynamic regulatory networks that rewire at different stages is a reasonable way to understand the mechanisms underlying cancer development. In this study, we reconstruct the stage-specific gene regulatory networks (GRNs) for colorectal cancer to understand dynamic changes of gene regulations along different disease stages. We combined multiple sets of clinical transcriptomic data of colorectal cancer patients and employed a supervised approach to select initial gene set for network construction. We then developed a dynamical system-based optimization method to infer dynamic GRNs by incorporating mutual information-based network sparsification and a dynamic cascade technique into an ordinary differential equations model. Dynamic GRNs at four different stages of colorectal cancer were reconstructed and analyzed. Several important genes were revealed based on the rewiring of the reconstructed GRNs. Our study demonstrated that reconstructing dynamic GRNs based on clinical transcriptomic profiling allows us to detect the dynamic trend of gene regulation as well as reveal critical genes for cancer development which may be important candidates of master regulators for further experimental test.

Keywords: clinical transcriptomic data; colorectal cancer; dynamic gene regulatory network; reconstruction

1. Introduction

Colorectal cancer (CRC), also named as colon cancer, is among top 3 common cancer diseases worldwide. With the fast development of powerful bioinformatics tools like microarray and high-

throughput sequencing technique, the researchers are able to have a peek into the gene expression level of such disease. Omics data-based modeling and analysis does not only provide opportunity of systematically uncover the molecular mechanism underlying CRC development but also facilitate designing more effective treatments and prognosis.

Many studies focus on exploring the gene expression profile to select differentially expressed genes and connect their known gene function with the disease. In such analysis, genes are viewed as independent molecules and their connections are often ignored. As such, the conventional differential gene expression studies fall short on investigating key genes at systems level. Network-based methods have been used to analyze high-throughput transcriptomic data of cancer patients [1]. The prevalent methods of reconstructing gene networks from cross-sectional clinical transcriptomic data mainly include correlational methods, such as PCC-based methods [2], mutual information [3], regression methods [4] or machining-learning methods [5], which can only infer correlations or associations between genes. However, these network reconstruction methods cannot infer directions of the edges in the network which is important for detecting regulatory relationship between genes underlying disease development.

To reconstruct gene regulatory networks (GRNs) from gene expression data, many methods have also been developed in the past years [6]. The typical methods include Boolean network methods [7], ordinary differential equation (ODE) methods [8,9], Bayesian network methods [10]. These methods usually require time-course gene expression data [11] for inferring directed GRN. However, the temporal information is always not available for the clinical transcriptomic data of patients, which restricted to apply these methods.

On the other hand, many studies employed static network approach to bridge the connection between molecular profile and biological functions [1]. However, cancer is a dynamic evolving disease and the intrinsic molecular interactions should also have dynamic changing patterns. Therefore, constructing dynamic regulatory networks that are rewiring at different stages is a rational way to understand the mechanism underlying cancer development.

In this paper, to detect dynamic rewiring of gene regulation network along different disease stage of colorectal cancer, we reconstruct the stage-specific gene regulatory networks from a combined colorectal cancer dataset. We first used a supervised approach to select genes for network construction and analysis by integrating genes' prognostic significance, functional association and correlation relationship and dynamic trends. An initial gene network was constructed using mutual information, which was then refined by employing a dynamic system model-based optimization approach to reconstruct the dynamic GRNs based on different colorectal cancer stages. The dynamic changes of the GRNs were analyzed to reveal important genes and their regulations during CRC development.

2. Materials and methods

2.1. Data collection and integration

In this article, four datasets of different sources are collected from both local hospital and gene expression omnibus (GEO) repository. GSE12945 is collected from Germany Max Planck Institute [12]; GSE14333 is collected from Australia Melbourne Royal Hospital [13] and GSE17536 is collected from United States Moffitt Cancer Center [14]. We also used a dataset of a Chinese cohort in our previous study, where 67 colorectal samples are processed through microarray analysis [15]. Both the gene

expression data and clinical information of patients were collected.

After close examination of the datasets, we found that the sample number of different cancer grade is unequally distributed among the four datasets, in which GSE12945 does not include grade A and grade D samples while GSE14333 does not include grade D samples. The best way to solve the inequality of sample number is to combine them together as a largely merged dataset, where the sample number of each grade is more balanced. Using selected variables including GSE IDs, overall survival months, living status, gender, cancer grade, age at diagnosis and data source, we combined the 4 different datasets together, resulting in 529 samples with valid phenotype information. Detailed information can be found in Table 1. We used 'removeBatchEffect' in limma R package to normalize the merged data and to reduce the between-experiments batch effects, which is used in the downstream analysis.

Table 1. Detailed information about four colorectal cancer datasets and merged dataset.

	Chinese cohort	Miffitt (GSE17536)	Max Planck (GSE12945)	Melbourne (GSE14333)	Merge (no NAs)
Sample (n)	67	177	62	226	529
Male, n (%)	41 (61)	96 (54)	34 (55)	120	288 (54.4)
Female, n (%)	26 (39)	81 (46)	28 (45)	106	241 (45.6)
Age (years)					
Median	59	66	—	67	66
Range	19–92	26–92	—	26–92	19–92
Cancer Grade					
A	4	24	0	41	61
B	16	57	31	94	276
C	29	57	31	91	177
D	18	39	0	0	15
Deaths	19	73	12	176	262
Median Survival Time (days)	1617.5	1268.1	1395	1153.8	1161.6

2.2. Mutual information estimation

Mutual information can be used to measure the nonlinear correlation relationship between two genes, which is a bilateral relation measurement. We use $I(A, B)$ to denote the mutual information between gene A and gene B. If $I(A, B) = 0$, then two gene expression patterns are uncorrelated, while larger $I(A, B)$ stands for closer pattern similarity and potentially implies close biological relationship.

In order to construct a mutual information network, we first calculate the mutual information value for each pair of genes, then a threshold is set to construct the connection network. In this step, the estimation method of mutual information plays a pivotal role in the performance of the ultimate mutual information network. Currently, two methods are commonly used to estimate the mutual information. First is to increase the accuracy of the estimation of joint probability density function from probability density estimation, such as histogram method and kernel density method [16]. The second way is to avoid the estimation of joint probability density function, such as K-nearest neighbor (KNN) method [17].

In terms of gene expression data, kernel density estimation is a better way because of its advantage

over the high-dimensional dataset and the intrinsic normal distribution pattern of gene expression values. Define $X_1, X_2, \cdots, X_n$ as independent identically distributed random variables with joint probability density function $f(x)$, then the kernel density estimation for function $f(x)$ is

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right), x \in R \tag{1}$$

where $K(\cdot)$ is called kernel function, and h is a pre-defined positive number, often referred as bandwidth funcion.

From the calculation of mutual information above, we know that the estimation for mutual information $I(X,Y)$ is $\hat{I}(X,Y)$, and it reads

$$\hat{I}(X,Y) = \iint_{X,Y} \hat{f}(x,y) log \frac{\hat{f}(x,y)}{\hat{f}(x)\hat{f}(y)} dxdy. \tag{2}$$

2.3. Network sparsification

Based on the above MI estimation of each pair of genes, we apply data process inequality (DPI) algorithm [18] to eliminate indirect associations between genes. In this way, a sparse network could be constructed, which is more suitable for the following regulatory network estimation.

Given the following gene interactions: $X \to Y \to Z$, where Z is correlated with Y but not with X, then the DPI algorithm considers that

$$I(X,Y) \geq I(X,Z) \text{ and } I(Y,Z) \geq I(X,Z) \tag{3}$$

That is, if (X, Y) and (Y, Z) are correlated directly but (X, Z) is indirectedly correlated through Y, then $I(X,Z) \leq I(X,Y)$, and $I(X,Z) \leq I(Y,Z)$. Using the DPI algorithm, we could exclude those indirect correlations, and retain those estimated to be directly associated gene pairs. We use e_{ij} to denote the edges in the MI network after network sparsification using DPI algorithm. $e_{ij} = 1$ if there is a direct association between gene *i* and gene *j*, otherwise $e_{ij} = 0$.

2.4. Dynamic GRNs reconstruction

Assume $x_i(t)$ to be the expression level of gene i at time t, then the ordinary differetial equation (ODE) of transcriptionl kinetics would be

$$\frac{dx_i(t)}{dt} = \sum_{j \in R_i} e_{ij} \cdot \beta_{ij} \cdot x_j(t) - \alpha_i \cdot x_i(t) \tag{4}$$

where α_i is the mRNA turnover rate, R_i is the set of regulators of gene i and β_{ij} is the regularotry strength from gene j to gene i. We incorporate e_{ij} derived from the MI network into the above ODE model to exclude indirectly-correlated gene pairs from the network model for the purpose of reducing computational cost in the following parameter estimation steps. If the time-course data are provided, this equation can be utilized to construct a dynamic gene regulation networks (GRN), and the coefficients can be estimated by a linear regression method. The coefficients can also be used to demonstrate the gene interaction in a GRN.

Many available cancer datasets for research use is stage-static data, in which all the case with the same cancer progression grade will be counted without time-course data. Thus, it makes people hard to examine the dynamic gene evolving process. In order to overcome such obstacle, we employed a dynamic cascaded methods (DCM) algorithm [19] (Text S1) to reconstruct the dynamic gene networks from sample-based transcriptional data.

According to the intra-stage steady-rate assumption and the continuity assumption [19], we are able to build up a model that connects two consecutive stages in terms of gene profile.

$$\bar{x}_i^{(s)} = -a_i^{(s-1,s)} \cdot \bar{x}_i^{(s-1)} + \sum_{j \in R_i^{(s-1)}} (b_{ij}{}^{(s-1,s)} \cdot \bar{x}_j^{(s-1)}) + \sum_{j \in R_i} (b_{ij}{}^{(s)} \cdot \bar{x}_j^{(s)}) \tag{5}$$

where $\bar{x}_i^{(s)}$ and $\bar{x}_i^{(s-1)}$ stand for the mean expressions of gene i in stage s and $s-1$. In addition, the coefficients in this formula is closely related to the previous ODE formula:

$$a_i^{(s-1,s)} = \frac{2-\alpha_i L^{(s-1)}}{2+\alpha_i L^{(s)}},\ b_{ij}{}^{(s-1,s)} = \frac{L^{(s-1)}}{2+\alpha_i L^{(s)}} \cdot e_{ij} \cdot \beta_{ij}^{(s-1)}, b_{ij}{}^{s} = \frac{L^{(s)}}{2+\alpha_i L^{(s)}} \cdot e_{ij} \cdot \beta_{ij}^{(s)} \tag{6}$$

where $L^{(s)}$ is the time length of stage s, α_i is the degradation rate, and $\beta_{ij}^{(s)}$ is the regulatory coefficient.

In order to describe the dynamic pattern within a stage, we should define the λ fraction of a stage as the proportional interpolation between the earliest and latest time points of the stage. According to the previous two assumptions, gene expression should be of linear trend change continuously, hence the largest and smallest gene expression should be on the either end of a gene expression stage. $t_1^{(s)}$ and $t_{N^{(s)}}^{(s)}$ denotes the earliest and the latest time points of stage s. The time of the λ fraction in stage s can be expressed as $t_\lambda^{(s)} = t_1^{(s)} + \lambda \cdot \left(t_{N^{(s)}}^{(s)} - t_1^{(s)}\right)$. After certain transformations, the dynamic model equation can be written as:

$$\begin{aligned} x_i^{(s)}\left(t_\lambda^{(s)}\right) = -a_i^{(s-1,s)} \cdot x_i^{(s-1)}\left(t_\lambda^{(s-1)}\right) + \sum_{j \in R_i^{(s-1)}} (b_{ij}{}^{(s-1,s)} \cdot x_j^{(s-1)}\left(t_\lambda^{(s-1)}\right)) \\ + \sum_{j \in R_i^{(s)}} (b_{ij}{}^{(s)} \cdot x_j^{(s)}\left(t_\lambda^{(s)}\right)) \end{aligned} \tag{7}$$

In this equation, gene expressions at the same (λ) fraction of two consecutive stages are connected. It describes the inter-stage dynamics of the GRN. And the intra-stage dynamical GRN can be described by the former ODE equation.

The details of the DCM algorithm are described in Text S1. All the steps listed above are performed in R (version 3.3.3) using packages including parallel, annotate, hgu133a.db, igraph, and most importantly, glmnet package that allows to fast implementation of elasticnet regression in selecting appropriate penalty coefficient and equation coefficients.

3. Results

3.1. Description of the merged datasets

In the merged dataset, there are about an equal number of male and female samples, counting for 54.4 and 45.6% respectively. The median age in those patients is 66 years old, corresponding to the higher hazard rate for old people to suffer from colorectal cancer. There are also young patients, who are only 19 years old, while the oldest patients are in their 90s. When it comes to the number of samples for each cancer grade, Grade B has the largest group of samples, while Grade D has the lowest group of samples, which is only 15. This phenomenon is also reasonable considering the terrible prognosis and lack of proper treatment for colorectal cancer in its late days. Throughout the whole survey, there are in total 262 recorded death cases and the median survival time length by day unit is 1161.1.

3.2. Gene set selection and enrichment

To reconstruct dynamic GRNs during CRC progression, we focus on cancer progression-related genes that should be functionally important and associated to the survival of patients. So, we used Cox PH model in this step to select the survival-associated genes for downstream analysis. The log-rank test p-value was calculated to assess the significance of the association between a given gene and survival of patients. We selected 1868 genes with p-value less than 0.01 for the downstream analysis, to reduce probability of Type I error.

We next used Gene Ontology to perform gene set enrichment analysis [20]. The most common method for GO enrichment analysis is the Fisher Exact Test, where contingency tables are made for each GO terms about the number of genes have or don't have the annotation and whether or not the genes are considered significant. 1868 survival-associated genes (log-rank p value < 0.01) were subject to gene set enrichment analysis. 'topGO' R package was used to perform the GO enrichment analysis and the top 10 GO functions are selected in Table 2, ranked by Kolmogorov-Smirnov test with more conservative 'elim' algorithm, which will render a more confident enrichment analysis result. From Table 2 we can see that the most enriched GO terms are closely related to the regulation of ubiquitin-protein, cell division, DNA elongation and replication and protein transportation, which have been verified to be closely related to cancer progression. For example, the deregulated ubiquitin ligases (E3s) have been experimentally validated to play important roles in tumor progression and cancer drug resistance [21]. For another example, protein trafficking has been shown to facilitate invasion and metastasis of CRC cells, and impaired transportation of member receptors represents important mechanisms of resistance to CRC-targeted therapy [22]. Furthermore, cell division and DNA elongation and replication are well-known biological processes associated with progression of many cancers including CRC [23].

To select genes with known direct interaction or indirect associations and narrow the range to select both statistically and biologically significant genes in colorectal cancer, we used STRING database (https://string-db.org/) to construct a PPI network using 1868 genes identified in the previous step as input. The PPI link tsv file is downloaded and 249 unique genes are selected in consideration of potential biological meanings.

Because of the multiple Affymetrix IDs for the same gene symbol, we here have more than 249 genes selected out of the merged gene expression matrix. Eventually, a matrix used to plot heatmap is

selected with 539 genes as rows and 529 patients as columns. The heatmap (Figure 1**)** shows that these genes are largely correlated with living status, which is shown in the first row of column side colors (green for live and red for death). Many living patients are associated with the higher expression level of these 539 selected genes.

Table 2. GO Enrichment Analysis Result.

GO.ID	Annotated	Significant	Expected	KS test p value	Term
GO: 0051436	111	12	9.1	3.90×10^{-5}	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
GO: 0051437	120	13	9.83	8.80×10^{-5}	positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition
GO: 0051301	814	72	66.71	1.10×10^{-4}	cell division
GO: 0006273	13	4	1.07	1.30×10^{-4}	lagging strand elongation
GO: 0045740	104	18	8.52	1.30×10^{-4}	positive regulation of DNA replication
GO: 0030049	59	7	4.84	1.30×10^{-4}	muscle filament sliding
GO: 0035999	14	4	1.15	1.40×10^{-5}	tetrahydrofolate interconversion
GO: 0009157	14	3	1.15	1.60×10^{-5}	deoxyribonucleoside monophosphate biosynthetic process
GO: 0032596	12	3	0.98	1.80×10^{-5}	protein transport into membrane raft
GO: 0031145	111	12	9.1	3.90×10^{-5}	anaphase-promoting complex-dependent catabolic process

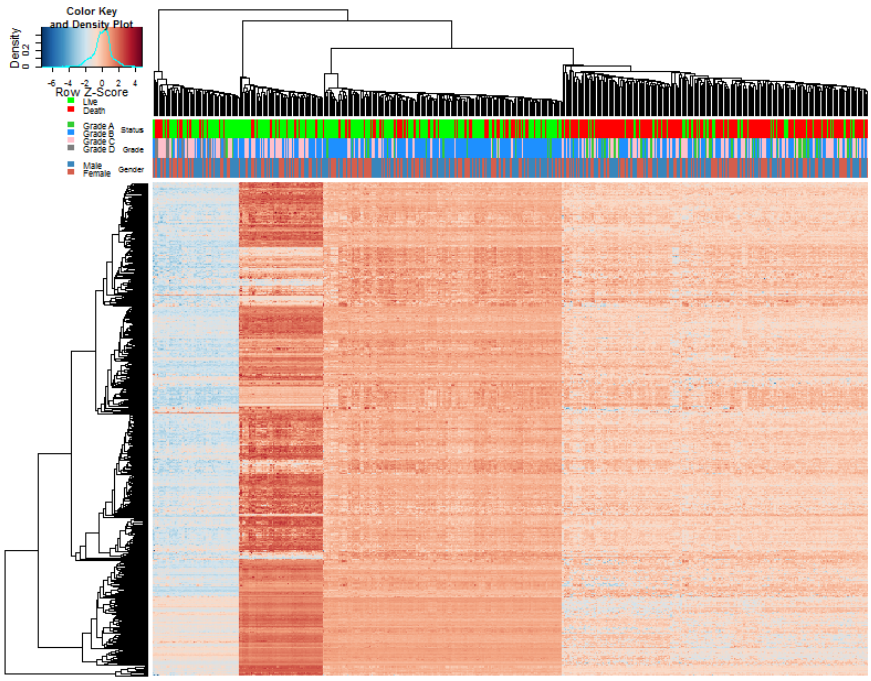

Figure 1. Heatmap plot of gene expression (539 genes of 529 patients).

3.3. Mutual information network

Using 249 genes selected from the previous STRING PPI network, we calculated mutual information for all pairs of genes and used the above-mentioned DPI algorithm to exclude indirectly associated gene pairs. In order to demonstrate the network structure, the Fruchterman-Reingold algorithm is used to automatically display the network as implemented in igraph R package. Because of the relatively large number of genes in our network, it's inappropriate to show them all together without a certain difference, hence we modify the text size, edge width and node size in the plot to help better visualize the weight of each gene. The text size is set to be proportionate to the ratio of node degree over maximum node degree of the network; edge width is set to be proportionate to the ratio of log transformed edge weight (which is also the mutual information value between gene pairs) over the log transformed maximum edge weight; and the node size is set to be the ratio of the sum of the mutual information value between one gene and all the other genes, over the maximum number of it.

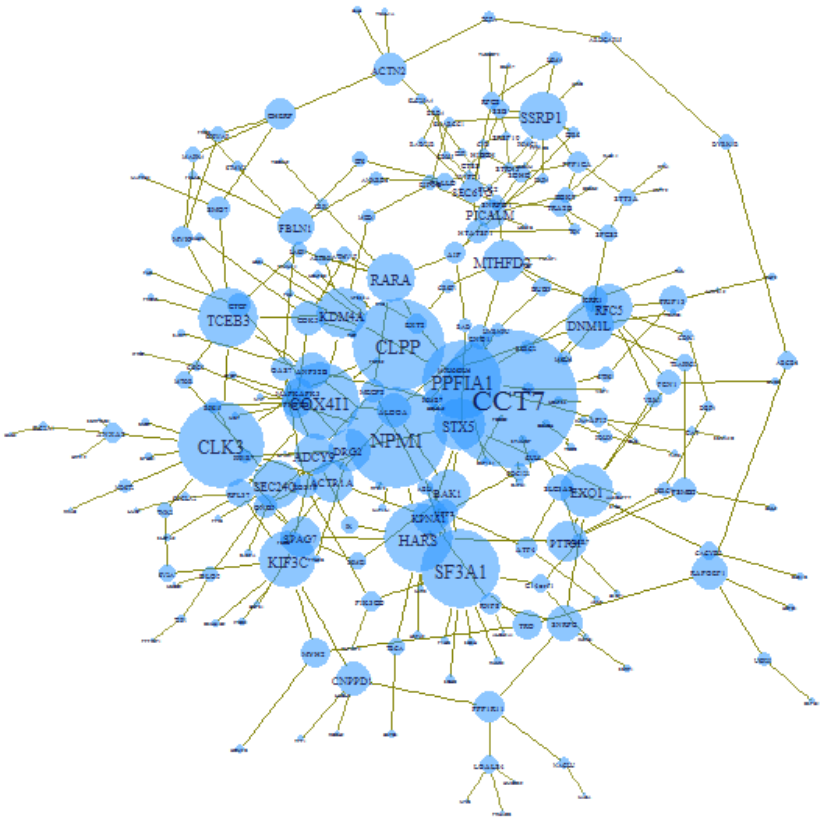

Figure 2. Mutual information network for the selected 249 genes.

The visualization of the constructed mutual information network is shown in Figure 2. We also examined the degree distribution (Figure 3a) and cumulative degree distribution (Figure 3b) for the mutual information network. Most genes (about 45%) have only one connection, verifying the sparsity of the constructed MI network using ARACNE method and DPI algorithm. High-degree genes in gene networks are often considered as hub genes or potential key genes which may possess important biological functions. In our analysis, we found that gene CCT7 does not only has the largest text size

but also it has the node size (Figure 2), indicating that CCT7 has both the largest connection with other genes and the sum of mutual information value. Importantly, CCT7 has been experimentally validated to be closely associated with some cancers, such as endometrial carcinoma [24]. This generates hypothesis of CRC as a key gene in CRC for experimental tests.

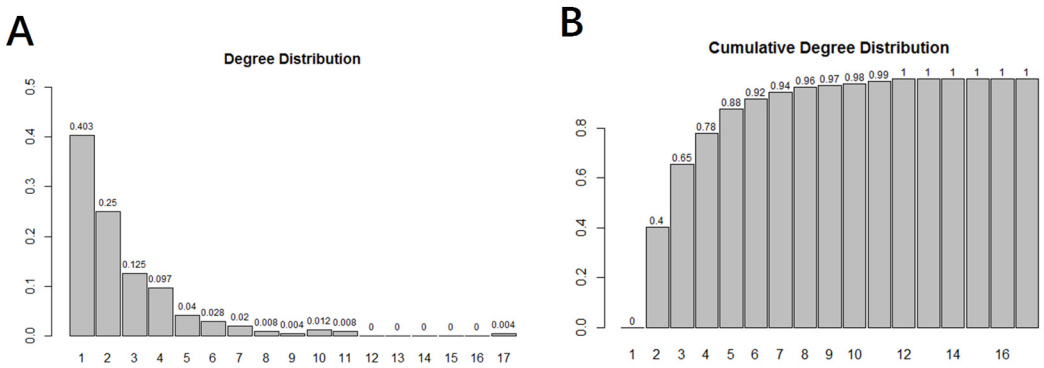

Figure 3. Node degree distribution of mutual information network. (A) Separate degree distribution, (B) cumulative degree distribution.

3.4. Reconstruction of dynamic GRNs

The mutual information network enables us to have a big picture of the gene association within CRC, however, it cannot provide more information about the association direction or the level of such association. To investigate the gene regulatory roadmap underlying CRC, we next infer the directed regulatory networks for each stage of CRC.

A smaller gene expression matrix (with 127 Affymetrix probes and 529 patients) is extracted from the original matrix (with 22277 Affymetrix probes and 529 patients). The genes are selected based on the following criteria: Survival association significance assessed (genes with log-rank test p-value less than 0.01), PPI interaction information (249 genes in String database) and mutual information network information (genes with the sum of mutual information > 1). The column number of the gene expression matrix corresponds with the phenotype dataset (which has 529 patient entries). Sample-based gene-profiling data of different colorectal cancer grade is created with 61 Grade A, 276 Grade B, 177 Grade C and 15 Grade D patients. The difference in the cancer grade represents the stepwise cancer progression process from early stage to late stage.

To determine the gene expression evolving direction, a gene evolving trend analysis was conducted. The overall connecting error is used to help to solve this problem. It is defined as the L1-norm of all the individual connecting errors:

$$\sum_{S=2}^{S} \left| x\left(t_{N^{(s-1)}}^{(s-1)}\right) - x\left(t_1^{(s)}\right) \right| \tag{8}$$

where $t_1^{(s)}$ and $t_{N^{(s-1)}}^{(s-1)}$ are the starting times of stage s and the ending time of stage s-1 respectively, $x\left(t_1^{(s)}\right)$ and $x\left(t_{N^{(s-1)}}^{(s-1)}\right)$ are the corresponding gene expression level of one gene. According to the continuity assumption, $x\left(t_1^{(s)}\right)$ and $x\left(t_{N^{(s-1)}}^{(s-1)}\right)$ should be the same. However, in practice, there

could be error at the connecting point of two stages. By summing up the overall connecting error, we can minimize it in order to get the shortest path from grade A to grade D. Then, this path is used to determine the gene evolving trend within each stage. Figure S2 illustrates the possible combinations between each stage.

The gene trend analysis is realized in R by using igraph package. First, the grade separated gene-profiling data is generated as large list object, in which every list in the root is the expression profile, and four sublists contain different gene expressions in different cancer stages. Then, quantile function is used to select genes according to a different λ fraction in a stage. λ is selected to be 0, 0.5, and 1 for interpolation. As for extrapolation, λ_{head} and λ_{tail} is set to be both at 4 for better controlling the linearity of the model equations. After that, a bootstrapping method is used to exploit the limited samples. We heuristically set the bootstrap sample size to half of the selected genes, and a total of 6150 bootstrap groups are generated for all the genes, and a total of 30750 corresponding model equations are obtained. The model coefficients in Eq (7) are then estimated by a regularized regression method. In this study we used elastic-net method to guarantee both sparsity and robustness of the estimation. We formulated the following objective function with a penalty to estimate ${b_{ij}}^{(s)}$ for each s,

$$\min\left\{\left\|x_i^{(s)}\left(t_\lambda^{(s)}\right)+a_i^{(s-1,s)}\cdot x_i^{(s-1)}\left(t_\lambda^{(s-1)}\right)-\sum_{j\in R_i{}^{(s-1)}}\left({b_{ij}}^{(s-1,s)}\cdot x_j^{(s-1)}\left(t_\lambda^{(s-1)}\right)\right)-\sum_{j\in R_i{}^{(s)}}\left({b_{ij}}^{(s)}\cdot x_j^{(s)}\left(t_\lambda^{(s)}\right)\right)\right\|_2^2+\ \theta\cdot\left[\frac{1-\alpha_{ela}}{2}\cdot P_1+\alpha_{ela}\cdot P_2\right]\right\},\ \mathrm{s}=2,3,4. \tag{9}$$

where α_{ela} is an elasticnet mixing parameter (between 0 and 1), and θ is the penalty coefficient. $P_1=\sum_i\left(a_i^{(s-1,s)}\right)^2+\ \sum_{ij}\left({b_{ij}}^{(s-1,s)}\right)^2+\ \sum_{ij}\left({b_{ij}}^{(s)}\right)^2$, and $P_2=\sum_i\left|a_i^{(s-1,s)}\right|+\sum_{ij}\left|{b_{ij}}^{(s-1,s)}\right|+\ \sum_{ij}\left|{b_{ij}}^{(s)}\right|$.

Cross validation strategy is also used to select the optimal θ. After getting the ${b_{ij}}^{(s)}$ matrix, we then convert them to $\beta_{ij}^{(s)}$ matrix according to previous formula Eq (6). Then a gene regulatory network (GRN) is constructed according to these $\beta_{ij}^{(s)}$. The network construction process is repeated 50 times, and ultimately 50 networks are obtained. The confidence of a connection is calculated as occurring frequency among 50 networks, and network connectivity P_{ER} is defined as the number of connections reserved in a network.

A sequence of α_{ela} is selected in our model to optimize the sparse gene regulatory network. Confidence-Connectivity plot (C-C plot) is used to demonstrate the relationship between the threshold confidence level of a network and the network connectivity. From the Figure S3, we found that the GRNs of Grade A and C have a similar curve, while Grade B is a little higher in the network confidence level as the network connectivity increases. Grade D curve has a similar descending curve to Grade A and C in the beginning, but then diverts from all the other 3 stages and descends more slowly. Among 15006 possible connections (with direction) for a set of 123 genes in colorectal cancer, most connections have relatively low confidence, in other words, low frequency of occurrence. The connection confidence decreases following an exponential trend as the network connectivity increases.

By setting a cutoff value of 30% confidence level, there is no more than 40% connectivity in either grade of the GRN (2615, 4999, 3157 and 4542 connections in Grade A, B, C, and D respectively).

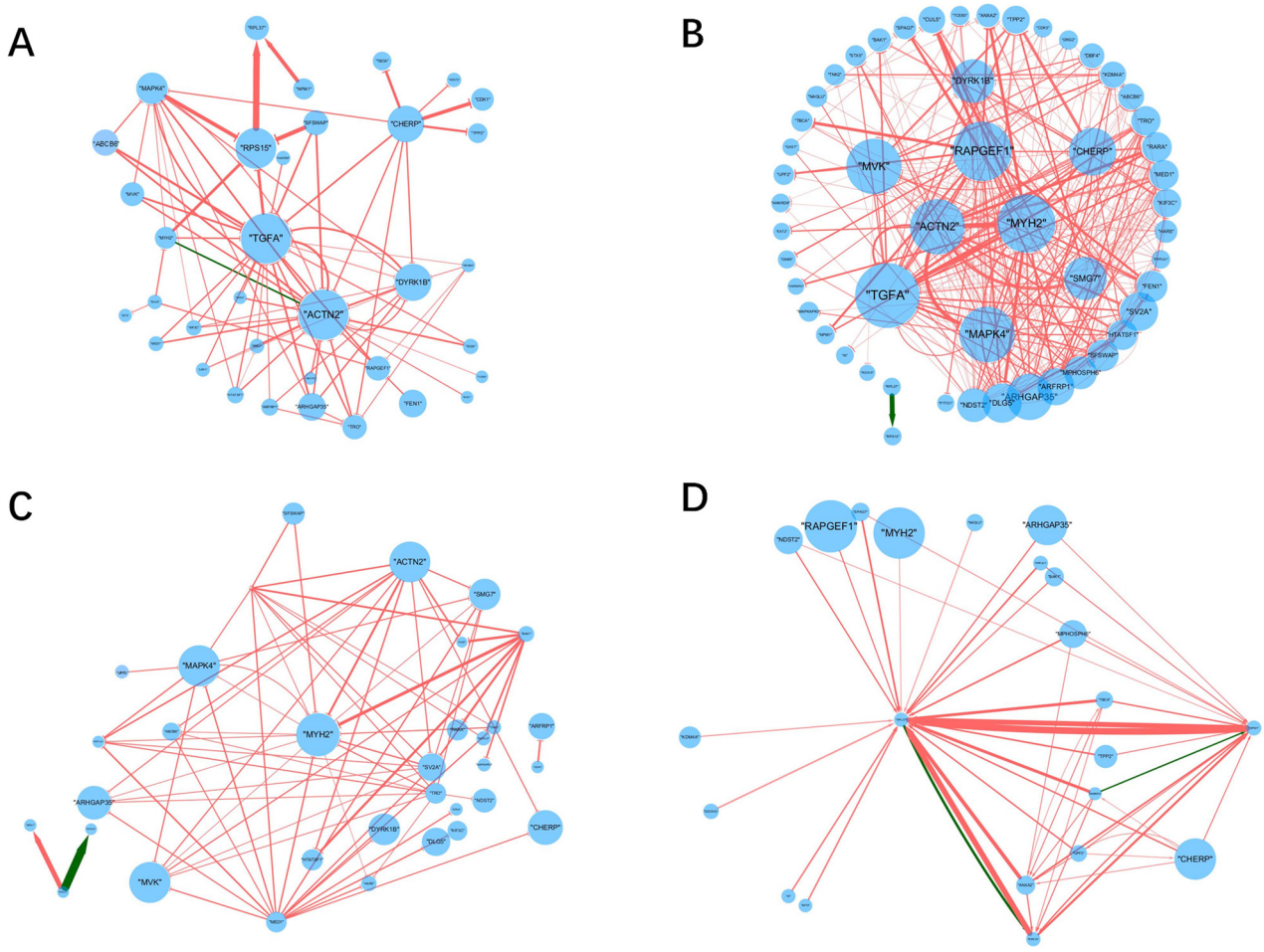

Figure 4. Reconstructed dynamic GRNs for four stages of CRC.

Four stage-specific gene regulatory network (GRN) are reconstructed with the connectivity setting of 0.02. Four GRNs are then used by Cytoscape (http://www.cytoscape.org) and shown in Figure 4. In these figures, the node size is set to be proportionate to the sum of the products of the confidence and strength over all the direct incoming and outgoing connections. This product stands for the activity of a gene in the network. Whether the strength is for upregulation and downregulation of the target gene is marked with either an arrow (→) or a stop (⊣). The level of the orange color of an edge is proportionate to the confidence level of the connection, and the width of the line is set corresponding to the strength of this regulation. The circular layout is ordered by the degree of the nodes.

3.5. Analysis of dynamic GRNs

In the following, we discuss the biological implications of the reconstructed dynamic GRNs for CRC at each stage.

During CRC stage A (Figure 4a), gene RPS15 has a big positive regulation over RPL37, and the former is negatively regulated by SFSWAP, MAPK4 and other two genes. One known connection with

high confidence level is the negative regulation from MYH2 to ACTN2.

During stage B (Figure 4b), RAPGEF1, ACTN2, MYH2, TGFA, MAPK4, SMG7, CHERP and DYRK1B are potential genes with biological significance. On literature review, RAPGEF1 gene have its first intron under somatic demethylation of a relaxed-criterion CpG island in 40% of colon cancers [25]. ACTN2, as isoform of ACTN4, whose amplifications has been verified to be associated with worse outcomes of cancer patients [26], is also believe to be critical to the outcome of CRC patients. MYH2 was found to be highly up-regulated (with 63-fold change) in CRC patients with tumors of high intrinsic COX-2 expression. TGFA is of the same family of TGF-beta, whose dysregulation has been shown to be closely related to colorectal cancer [27]. CHERP was found in Colorectal Cancer Atlas [28] with functions of negative regulation on cell proliferation.

In CRC stage C (Figure 4c), there is an interesting pattern of negative regulation from gene BAK1 to several other genes such as RAPGEF1, MYH2. Although BAK1 has relatively small value of product sum (corresponding to gene activity), it has a broad regulation function on multiple genes. Another interesting pattern to notice is that genes with low activity might have a strong regulation effect on other genes, such as RPL37. This is reasonable because important regulatory genes are not always active while their effects could be tremendous.

Figure 4d shows the stage D-specific regulatory network. The previous observation reoccurred that two low-expressed genes RPL37 and NPM1 have strong influence on the expression of each other as well as being regulated by many other active genes. From this pattern we hypothesize that those low-expressed but influential genes might be the downstream gene regulated by multiple less influential but high-expressed genes. The additive effect of the upstream genes might largely affect the downstream biological process.

4. Discussion

In this study, we employed an integrated approach to analyze clinical transcriptomic data of CRC patients. We are interested at the directional gene regulatory networks for each stage and their dynamic changes during CRC development. Driven by this question, we collected and merged 4 datasets of CRC containing both gene expression data and clinical information. To infer GRNs from the large-scale clinical gene expression dataset, the survival-associated genes were selected as input for network construction, and an initial mutual information network was constructed to reduce the computational cost of the parameter estimation of the dynamical system model of GRN. By introducing a new regularized optimization method into the DCM algorithm, directional GRNs were inferred for each stage of CRC.

The DCM algorithm [19] employed in this paper relies on two assumptions: The intra-stage steady-rate assumption and the continuity assumption. The intra-stage steady-rate assumption or linear-dynamic assumption, models the dynamic gene expression profile with a linear trend within each stage of a process. The continuity assumption assumes that the gene expression trajectory is continuous both within each disease stage and at the connecting point of two adjacent stages. Obviously, these two assumptions are simplification of the real gene expression in patients. In reality, it's largely plausible that the gene expression dynamics should be nonlinear, and the gene mutations or other genetic variations may result in discrete transition of gene expression trajectory. Therefore, the linear-dynamic assumption and the continuity assumption need to be generalized or waived to be more realistic.

In addition, the number of samples in each stage (stage A to D) is uneven in our study. The minimal

sample number in that study is 13, which is comparative to 15 in our paper. Although the previous study has shown that the DCM algorithm is not sensitive to the sample size (Figure 3B in [19]), the uneven sample number could be considered for improving the DCM algorithm. For example, this consideration may be addressed by assigning different penalties or weights for different groups according to the number of samples.

In the current study, we focus on GRNs rewiring during CRC development from stage A to stage D. We did not consider normal samples of CRC since they are not available in the collected dataset, which is another limitation of our study. From the viewpoint of dynamic theory, disease progression can be considered a state transition, from a normal state, to pre-disease state and then to disease state. So, integration of normal samples into dynamic network analysis would reveal some important insights for cancer initiation and occurrence, which will be studied in our future study.

Inferring gene regulatory network from high throughput data is a fundamental question in systems biology yet remains as a challenging task. Analyzing GRNs and their dynamic rewiring are instrumental to identifying master regulator genes or casual drivers in human disease at systems level. The existing methods for inferring GRNs largely depend on time-course gene expression data. However, large samples of time-course data are rarely available in clinical situations since longitudinal surveys are often challenging to conduct. In contrast, cross-sectional studies of a large population based on high-throughput molecular omics data are more prevalent due to their relative feasibility.

The lack of temporal information in clinical transcriptomic data leads to a major challenge for inferring GRN and its translation to precision medicine. Therefore, how to decode the temporal information underlying disease progression is an interesting question and may be key to address the above challenge. The sample similarity-based approach in evolution and genetics studies [29], for instance, phylogenetic trees based on microarray data [30] and genetic linkage maps based on genetic markers [31], have shown developed and shown success in recovering evolutionary dynamics. Importantly, some studies have shown that it is possible to measure cancer progression by using gene expression profiles to estimate "distances" between tumors [32]. Another study used graph-based minimum spanning tree to recover biological progression from microarray samples [33], which can be applied for reconstructing temporal trajectory of differentiation, development, cell cycle or disease progression.

In summary, in this study, we analyzed the dynamic changes of GRNs along different stages of CRC by using an integrated approach. We combined several clinical transcriptomic datasets of CRC and reconstructed GRNs of CRC across different stages by integrating gene selection, mutual information-based network sparcification and a previous DCM algorithm. Our study demonstrated that the dynamic rewiring of gene regulations could be inferred by reconstructing dynamic GRNs based on clinical transcriptomic profiling. Our study revealed several key genes underlying CRC development, which may generate new hypothesis for further experimental test. We anticipate new methods in future studies to quantitatively estimate disease progression based on the clinical cross-sectional transcriptomic data, which may be promising for developing dynamical system-based GRN inference approach.

Data availability

The gene expression datasets as well as clinical information of the patients were downloaded from the NCBI GEO database, as described in the main text.

Acknowledgements

X Sun was supported by grants from the National Key R&D Program of China (2018YFC0910500), the National Natural Science Foundation of China (11871070), the Guangdong Basic and Applied Basic Research Foundation (2020B1515020047), Guangdong Key Field R&D Plan (2019B020228001) and the Opening Project of Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (2018003).

Conflict of Interests

The authors declare that they have no conflict of interest.

Author contributions

X.S. designed research; A.D. performed research; A.D. and X.S. analyzed data and wrote the paper.

References

1. X. Sun, B. Hu, Mathematical modeling and computational prediction of cancer drug resistance, *Briefings Bioinf.*, **19** (2018), 1382–1399.
2. B. H. Liu, Differential Coexpression Network Analysis for Gene Expression Data, in *Computational Systems Biology: Methods and Protocols* (ed T. Huang), Springer, New York, (2018), 155–165.
3. K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, A. Califano, Reverse engineering of regulatory networks in human B cells, *Nat. Genet.*, **37** (2005), 382–390.
4. R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, *J. R. Stat. Soc.*, **58** (1996), 267–288.
5. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring Regulatory Networks from Expression Data Using Tree-Based Methods, *PLOS ONE*, **5** (2010), e12776.
6. N. Le Novère, Quantitative and logic modelling of molecular and gene networks, *Nat. Rev. Genet.*, **16** (2015), 146–158.
7. C. H. A. Higa, T. P. Andrade, R.F. Hashimoto, Growing Seed Genes from Time Series Data and Thresholded Boolean Networks with Perturbation, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2013. **10** (2013), 37–49.
8. S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, M. Tomita, Dynamic modeling of genetic networks using genetic algorithm and S-system, *Bioinformatics*, **19** (2003), 643–650.
9. Zhang, J., et al., Differential regulatory network-based quantification and prioritization of key genes underlying cancer drug resistance based on time-course RNA-seq data, *PLOS Comput. Biol*., **15** (2019), e1007435.
10. M. Grzegorczyk, D. Husmeier, K. D. Edwards, P. Ghazal, A. J. Millar, Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler, *Bioinformatics*, **24** (2008), 2071–2078.
11. Y. Kim, S. Han, S. Choi, D. Hwang., Inference of dynamic networks using time-course data, *Briefings Bioinf.*, **15** (2014), 212–228.

12. E. Staub, J. Groene, M. Heinze, D. Mennerich, S. Roepcke, I. Klaman, et al., An expression module of WIPF1-coexpressed genes identifies patients with favorable prognosis in three tumor types, *J. Mol. Med.*, **87** (2009), 633–644.
13. R. N. Jorissen, P. Gibbs, M. Christie, S. Prakash, L. Lipton, J. Desai, et al., Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer, *Clin. Cancer Res.*, **15** (2009), 7642–7651.
14. J.. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, A. Jiang, et al., Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer, *Gastroenterology*, **138** (2010), 958–968.
15. H. Chen, X. Sun, W. Ge, Y. Qian, R. Bai, S. Zheng, A seven-gene signature predicts overall survival of patients with colorectal cancer, *Oncotarget*, **8** (2017), 95054–95065.
16. Y. I. Moon, B. Rajagopalan, U. Lall, Estimation of mutual information using kernel density estimators, *Phys. Rev.*, **52** (1995), 2318–2321.
17. W. M. Lord, J. Sun, E. M. Bollt, Geometric k-nearest neighbor estimation of entropy and mutual information, *Chaos*, **28** (2018), 033114.
18. T. M. Cover, J. A. Thomas, *Elements of information theory*, John Wiley & sons, New Jersey, 2003.
19. H. Zhu, R. S. P. Rao, T. Zeng, L. Chen, Reconstructing dynamic gene regulatory networks from sample-based transcriptional data, *Nucleic Acids Res.*, **40** (2012), 10657–10667.
20. The Gene Ontology Consortium., Gene Ontology Consortium: Going forward, *Nucleic Acids Res.*, **43** (2015), D1049–D1056.
21. D. Senft, J. Qi, Z. A. Ronai, Ubiquitin ligases in oncogenic transformation and cancer therapy, *Nat. Rev. Cancer*, **18** (2018), 69–88
22. A. N. Gargalionis, M. V. Karamouzis, C. Adamopoulos, A. G. Papavassiliou., Protein trafficking in colorectal carcinogenesis—targeting and bypassing resistance to currently applied treatments, *Carcinogenesis*, **36** (2015), 607–615.
23. M. J. Pillaire, J. Selves, K. Gordien, P. A. Gouraud, C. Gentil, M. Danjoux, A 'DNA replication' signature of progression and negative outcome in colorectal cancer, *Oncogene*, **29** (2010), 876–887.
24. N. Shan, W. Zhou, S. Zhang, Y. Zhang, Identification of HSPA8 as a candidate biomarker for endometrial carcinoma by using iTRAQ-based proteomic analysis, *Onco. Targets Ther.*, **9** (2016), 2169–2179.
25. J. Samuelsson, S. Alonso, T. Ruiz-Larroya, T. H. Cheung, Y. F. Wong, M. Perucho, Frequent somatic demethylation of RAPGEF1/C3G intronic sequences in gastrointestinal and gynecological cancer, *Int. J. Oncol.*, **38** (2011), 1575–1577.
26. K. Honda, The biological role of actinin-4 (ACTN4) in malignant phenotypes of cancer, *Cell Biosci.*, **5** (2015), 41.
27. A. Calon, E. Espinet, S. Palomo-Ponce, D. V. F. Tauriello, M. Iglesias, M. V. Céspedes, et al., Dependency of Colorectal Cancer on a TGF-β-Driven Program in Stromal Cells for Metastasis Initiation, *Cancer Cell*, **22** (2012), 571–584.
28. D. Chisanga, S. Keerthikumar, M. Pathan, D. Ariyaratne, H. Kalra, S. Boukouris, et al., Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues, *Nucleic Acids Res.*, **44** (2016), D969–D974.

29. D. L. Rabosky, M. Grundler, C. Anderson, P. Title, J. J. Shi, J. W. Brown, et al., BAMM tools: An R package for the analysis of evolutionary dynamics on phylogenetic trees, *Methods Ecol. Evol.*, **5** (2014), 701–707.
30. R. Desper, J. Khan, A. A. Schäffer, Tumor classification using phylogenetic methods on expression data, *J. Theor. Biol.*, **228** (2004), 477–496.
31. Y. Wu, P. R. Bhat, T. J. Close, S. Lonardi, Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph, PLoS Genet., **4** (2008), e1000212.
32. Y. Park, S. Shackney, R. Schwartz, Network-Based Inference of Cancer Progression from Microarray Data, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **6** (2009), 200–212.
33. Qiu, P., A. J. Gentles, S. K. Plevritis, Discovering Biological Progression Underlying Microarray Samples, *Plos Comput. Biol.*, **7** (2011), e1001123.

AIMS Press