[Abstract] Today’s fast developing biotechnology allows for high-throughput molecular-level research on organisms, especially on cancer tissues. Among all available techniques, microarray chip is one of the best options we can have to monitor the gene expression as well as build the bridge between expression and biological functions. In order to understand the gene regulation network along different disease stage, continuous time-course data is needed, however, often the situation is we can only access the expression profile from hospital, where samples are separated by their medical grading system. In this paper we are attempting to reconstruct the dynamic gene regulation network from a combined colorectal dataset using one method called DCM (dynamic cascaded method). The combined dataset is obtained from 4 sepearate sources: Vanderbilt University (GSE17536), Ludwig Institute for Cancer Research (GSE14333), Signature Diagnostics AG (GSE12945) and Chinese hospital (patient67). After data merging, Kaplan-Meier survival curve and Log-rank used to get a sense of the survival probability relative to different colorectal cancer stages and gender; then, Cox proportional hazard model (data source as risk factor) is chosen and selects 1868 significant genes with p-value threshold of 0.01; GO (gene ontology) analysis shows that these genes are mostly associated with regulation of ubiquitin-protein; then these 1868 selected genes are analyzed on their protein-protein interaction using online STRING analyzer, which further renders a sparsified gene interaction network and selects certain nodes with known connection with others. A network of 248 genes and 694 connections are kept after this process. Mutual information network is utilized to have a broad view of the correlation of expression value, and a filter is added to further diminish number of potential important genes based on the sum of mutual information score. Finally, a set of 127 genes is obtained, and the DCM method is implemented on our 127 genes set to reconstruct the dynamic GRN (gene regulatory network). Multiple attributes can be contained in this network, such as up/down regulation, strength of the regulation, confidence of a connection and the degree of nodes (genes). Reconstructing dynamic GRNs based on different colorectal cancer stages allows us to follow the dynamic trend of gene regulation as well as find novel target genes for medicine development and gene therapy.

[key words] Dynamic Cascaded Method; Dynamic Gene Regulatory Network; K-M curve; Cox Proportional Hazard Model

【摘要】 随着基因芯片技术的成熟和第二代测序技术的完善，研究人员能够更加方便地投入对任何有机物基因分子层面的研究当中去，其中以癌症的基因研究最为普遍。研究人员普遍希望探索某些在特定癌症中差异性表达的基因，并且跟踪这些基因在不同癌症阶段的表达水平变化，甚至是基因之间的调控。然而，由于数据的限制，目前研究人员绝大时候获得的是医院对于不同癌症阶段的不同病人的数据，一般是在病人进行检查时候取得，因此不能满足同一样本长时间取样的要求。在这种情况下，如何重组动态基因调控网络（GRM）成为了研究热点之一。本文采用了近些年用到的一种动态基因调控网络重组方法，在尽可能保持基因阶段性变化的前提下判断和预测基因之间的调控模式。为了尽可能地保留潜在最重要的一些基因，我们在构建该网络之前使用了一些常用的生存和网络构建方法对于基因进行筛选，首先通过Cox Proportional Hazard模型选择p值小于0.01的1868个基因，接着将基因输入到STRING蛋白质互作网络中筛除与其他基因没有任何已知关系的基因，之后，将剩余的248个基因通过互信息网络模型进一步根据互信息的权重，最后将得到的127个基因作为动态GRN的节点。重建动态网络的过程依赖于Elastic Net弹性模型对于基因调控系数的线性回归估计，并通过网络连通度-连通置信度曲线选择合适的网络连通度。网络富集度分析和多特征网络构建能够帮助我们更好地理解网络的结构和组成，并且在此基础上发掘有重要性的基因和基因间关系。

【关键词】 动态基因调控网络；差异表达、CoxPH模型；互信息网络；STRING互作网络；DCM方法

Contents

1. Backgrounds 1

2. Purpose & Significance 1

3. Methods 1

3.1. Gene Microarray Datasets Collection 1

3.2. Gene Microarray Datasets Combination 2

3.3. Kaplan-Meier Survival Curve 2

3.4. CoxPH Analysis 3

3.5. Mutual Information Network 3

3.5.1. Entropy 4

3.5.2. Mutual Information 4

3.5.3. Estimation of Mutual Information 5

3.5.4. Network Sparsification Method 5

3.6. Dynamic Cascaded Method to Reconstruct Dynamic gene networks ... 6

3.6.1. Assumptions 6

3.6.2. Gene Regulatory Network & Dynamic Cascaded Methods 6

3.6.3. DCM algorithm 7

4. Results 8

4.1. Description of the merged datasets 8

4.2. Kaplan-Meier survival curve 9

4.3. Using Cox Proportional Hazard Model to select significant gene 10

4.4. Gene Ontology (GO) Enrichment Analysis 10

4.5. Protein-protein Interaction (PPI) with STRING 11

4.6. Heatmap Visualization 12

4.7. Mutual Information Network 13

4.8. Reconstruction the Gene Regulatory Network 14

4.9. Network Enrichment Analysis 16

4.10. Functional Analysis 17

5. Discussion 20

Reference: 21

Acknowledgement 22

1. Backgrounds

Colorectal cancer (CRC), also named as colon cancer, is among top 3 common cancer disease worldwide and rank 2 as female common diseases. Although the advancing healthcare service has successfully controlled the mortality of CRC, it remains the 3rd most common cause of cancer-related mortality, accounting for approximately 600,000 deaths in 2008 worldwide[1]. The American Cancer Society estimates that in the United States for 2017, there will be 95520 new cases of colon cancer along with 39910 new cases of rectal cancer. In terms of the difference between male and female, the male has more incidence rate than female (2-3: 1). Developed countries such as Australia, New Zealand, those Europe and North America share the soaring incidence rate, while under-developed countries in Africa, South Asia and Mid-America have the lowest. Possible causes of such regional difference include the change of diet, Type II diabetes, lack of exercise, obesity, over-smoking, excessive drinking and so on. It’s crucially important to have an early diagnose of CRC, because, with proper medical treatment, the five-year survival rate can be 90%; while stage IV CRC patients can only have 15%.

Currently, the primary treatment for CRC is a surgical operation with neoadjuvant radiotherapy or/and adjuvant chemotherapy, according to the tumor location and the stage of the disease. In general, these methods do not meet with people’s expectation and still undesirable prognosis. With the emergence of powerful bioinformatics tools like microarray and high-throughput sequence technique, people are allowed to have a peek into the gene expression level of such disease and locate the key to regulate the cancer progression. It will not only allow us to unveil the mechanism behind how the CRC occurs and develops but also it will provide a guide for the researches on more effective CRC treatments and prognosis.

2. Purpose & Significance

Currently, many types of research focus on exploring the gene expression microarray chip to select differentially expressed genes and connect their known gene function with the disease. Those selected genes may be crucial along the cancer progression steps such as transformation, dedifferentiation, vasculogenesis, tumor metastasis and tumor infiltration. Meanwhile, larger and larger datasets are being used to draw a more confident selection. In addition, tumor progression doesn’t rely on independent genes but a group of connected ones, which regulate others while be regulated. Gene regulation network is another hot spot these days to allow people to visualize and analyze the interaction between each gene. By reconstructing the the dynamic gene regulatory network and comparing the gene regulatory network in different cancer progression stages, we are more likely to find host gene that is crucial to a special disease.

3. Methods

3.1. Gene Microarray Datasets Collection

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[2]; GSE14333 is collected from Australia Melbourne Royal Hospital[3] and GSE17536 is collected from United States Moffitt Cancer Center[4]. We also collected data from Chinese hospital, where 67 colorectal samples are processed through microarray analysis.

For Chinese 67-patient dataset, we already have the log2 transformed gene expression matrix, which has 54675 rows (as Affymetrix IDs) and 67 columns (as patients); while for the other 3 datasets, the dataset is first collected using GEOquery package from R repository. The gene expression data along with the phenotype data (about patient information) are downloaded to local storage as series_matrix.txt files. Then, expression data and phenotype data are separated using Biobase package and string manipulation function in R to extract the GEO accession ID, tumor location, number of lymph node removed, living status, overall survival months, tumor free months, gender, age at diagnosis and colorectal cancer grade. Some datasets have all the information while others lack one or several variables. All the datasets along with the function used for data cleaning is stored in an RData object as original archive data. As for the Affymetrix IDs for these 3 datasets, GSE12945 has 22283 IDs while the other 2 have 54675 IDs. This ID number difference should be considered in the next step when gene expression datasets are to be merged.

3.2. Gene Microarray Datasets Combination

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 have no grade A and grade D samples while GSE14333 have no 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. After done with the combination process, in total 529 samples with valid phenotype information is kept and selected variables are GSE IDs, overall survival months, living status, gender, cancer grade, age at diagnosis and data source. Detailed information can be found in **Table 1**. Then the merged dataset is used in the downstream analysis.

3.3. Kaplan-Meier Survival Curve

Kaplan-Meier survival analysis is first brought up by Statistician Kaplan and Meier in 1958. This method uses conditional probability and the production principle to calculate survival rate and its standard error. Its most direct result is the K-M survival curve, which is widely used to track a group of objects (mostly patients) along time and records event-happening. One of the merits of a K-M curve is that it will take advantage of censored data (most commonly right censored ones) and hence take into account all available survival information. We might want to note that when considering censored dataset this method separate time region according to the actual happening events (such as actually death), so the censored data tends to have a higher curve relative to the uncensored. The survival probability at any particular time is calculated as below[5]:

$$\mathrm{S}_t = \frac{\#subjects\ living\ at\ the\ start - \#subjects\ died}{\#subjects\ living\ at\ the\ start}$$

Comparing survival curves is of particular interest in clinical trials. In addition to the visible difference in K-M curve, we also need to quantify the difference in order to estimate the statistical difference[6]. One common and effective method is log rank test, which calculates the chi-square (X^2) for each event time for each group and sums the results. The summed results are then added to get the ultimate X^2 so as to compare the full curves of each group. The power to reject the null hypothesis in this test is related to the sample size, which means that even though the curve is visually very different, statistically the power of the difference in the small group of the sample is not sufficient to rule out a real difference. In such case, we might need to worry about the type II error of the method we used.

$$\text{Log} - \text{rank test statistic} = \frac{(0_1 - E_1)^2}{E_1} + \frac{(0_2 - E_2)^2}{E_2}$$

3.4. CoxPH Analysis

Cox proportional hazard model is a semi-parameter model and can be regarded as a survival analysis with covariates: sometimes we want to analyze whether some potentially related factors (also known as independent variable or covariates) can be influential to the survival time, and the significance of their influence. Let’s assume **G** number of genes and **N** number of samples, in which we have observed overall survival time and living status. In the jth sample, Y_j and δ_j stand for the observed overall survival time and living status (0 for missing and 1 for death). $X_j = \left(X_{1j}, X_{2j}, \mathrm{K}\ , X_{Kj}\right)$ is the j th variable for K gene expression, where $K < N$ and $K \subset G$. $Y_{(1)} < Y_{(2)} < \mathrm{K}\ < Y_{(D)}$ represents ordered survival time with D different values, while $X_{(i)k}$ means the kth gene under the correlated sample $Y_{(i)}$.

The formula for Cox proportional hazard model is

$$h\left(y \left| X_1, X_2, \mathrm{K}\ , X_K \right.\right) = h_0\left(y\right) exp\left(\sum_{k=1}^{K} \beta_k X_k\right)$$

where $h\left(y \left| X_1, X_2, \mathrm{K}\ , X_K \right.\right)$ is the hazard ratio for the variable $\left(X_1, X_2, \mathrm{K}\ , X_K\right)$ under the time y, $h_0\left(y\right)$ is the baseline hazard function, and β_k is the parameter for the kth gene. The partial likelihood for CoxPH model is as follows:

$$\sum_{i=1}^{D} \sum_{k=1}^{K} \beta_k X_{(i)k} - \sum_{i=1}^{D} \ln\left[\sum_{j \in R\left(Y_{(i)}\right)} \exp\left(\sum_{k=1}^{K} \beta_k X_{jk}\right)\right]$$

where $R\left(Y_{(i)}\right)$ means all the training sample set in time point $Y_{(i)}$.

3.5. Mutual Information Network

Entropy and mutual information are an important compartment in Information Theory, which uses mathematical statistics methods to study the basic attributes of information and measures of it.

Information Theory is a type of science that commits to the best way to solve problems about information gathering, transmission, storage, process, and transformation. Mutual information is one measurement that considers the correlation between variables. Through association network constructed based on mutual information, we are better off to understand the connection between genes hence to better explore gene interactions.

3.5.1. Entropy

Entropy is a measure of uncertainty. Assume there are **n** mutually exclusive events A_1、A_2、 ⋯ 、A_n, one and only one of them will happen. Following a counting rule to number each event:

a) Events with equal probability will be assigned an equal number of symbol digits.

b) Events with large probability will be assigned with less number of symbol digits

In general, if we number the experiment results according to α-digit system, then event with probability α^{-m_i} will be assigned with $m_i(i = 1,2,\cdots,n)$ number of α-digit (m_i is natural number). The average number of symbol digits as follows:

$$H = \sum_{i=1}^{n} m_i \alpha^{-m_i}$$

H is the measurement for uncertainty for random events. For i^{th} event, its probability is $p_i = \alpha^{-m_i}(i = 1,2,\cdots,n)$, and former formula can also be written as:

$$H = \sum_{i=1}^{n} p_i \log_\alpha \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_\alpha p_i$$

And this is the measure of uncertainty for general finite discrete probability events.

3.5.2. Mutual Information

Mutual information can be such a value to measure the correlation between genes. Let's say the entropy for gene pattern A can be

$$H(A) = -\sum_{i=1}^{n} p(x_i)\log_2(P(x_i))$$

In which $P(x_i)$ is the frequency that gene expression value is found in the range of x_i, n is the number of ranges for gene expression level. A larger $H(A)$ usually means more randomly distributed gene expression. The mutual information between 2 gene expression pattern can be written as:

$$I(A,B) = H(A) + H(B) - H(A,B)$$

Mutual information is a bilateral relation measurement. If $I(A,B) = 0$, then two gene expression pattern is uncorrelated, while larger $I(A,B)$ stands for closer pattern similarity and maybe close biological relationship.

3.5.3. Estimation of Mutual Information

In order to construct a mutual information network, first is to calculate the mutual information value for every pair of gene expression pattern, 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[7] and kernel density method[8]. The second way is to avoid the estimation of joint probability density function, such as K-Nearest Neighbor (KNN) method.

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. Let's define $X_1, X_2, \cdots, X_n$ as iid random variables with joint probability density function $f(x)$, then we have

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

as the kernel density estimation for function $f(x)$. $K(\cdot)$ 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

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

3.5.4. Network Sparsification Method

Many network model built upon mutual information are now available to use, such as ARACNE, CLR, and MINET. Among these methods, ARACNE is one of the most important methods to sparsify the gene connection network. It applies DPI algorithm to the mutual information symmetric matrix consists of mutual information value between each pair of gene expression patterns.

Let's define the data process inequality (DPI). If there exists a Markov chain: $X \rightarrow Y \rightarrow Z$, where Z is correlated with Y but not with X, then:

$$I(X;Y|Z) = 0\text{，} \text{with } I(X,Y) \geq I(X,Z), I(Y,Z) \geq I(X,Z)$$

That's to say if gene x and gene y do not have direct correlation, instead of correlated through gene z, then conditional mutual information $I(x;y|z)$ is 0. In DPI algorithm, if (x, y) and (y, z) are correlated directly but (x, z) is indirectly correlated through y, then we believe the situation will be $I(x,z) \leq I(x,y)$, and $I(x,z) \leq I(y,z)$. When initiating the network building process, we use the DPI algorithm to select out those indirect correlations, and leave out those believed to be directly associated gene pairs, hence the network sparsification method.

3.6. Dynamic Cascaded Method to Reconstruct Dynamic gene networks

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 know and follow the dynamic gene evolving process. In order to overcome such obstacle, Dr. Hailong Zhu and his colleagues developed the dynamic cascaded methods (DCM) to reconstruct the dynamic gene networks from sample-based transcriptional data[9].

3.6.1. Assumptions

a) Intra-stage steady-rate assumption: it assumes that gene expression can be dynamic, and the dynamic profile should be associated with a linear trend within each stage of a process.
b) Continuity Assumption: it means that there are no discrete or abrupt changes in the gene profile even at the time of stage transition. It is relatively natural for the accumulated process like gene expression.

3.6.2. Gene Regulatory Network & Dynamic Cascaded Methods

Let’s 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:

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

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. Time course data 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.

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

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

in which $\bar{x}_i^{(s)}$ 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 correlated with the previous ODE formula:

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

in which $L^{(s)}$ is the time length of stage s, α_i is the degreadation rate and $\beta_{ij}^{(s)}$ is the regulatory strength.

In order to describe the dynamic pattern within a stage, we should define the $\lambda\,\text{fraction}$ of a stage as the proportional interpolation between the earliest and lastest 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 $\lambda\,\text{fraction}$ in stage s can be expressed as $t_\lambda^{(s)} = t_1^{(s)} + \lambda \cdot (t_{N^{(s)}}^{(s)} - t_1^{(s)})$. After certain transformations, the dynamic model equation can be written as:

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

$$+ \sum_{j\in {R_i}^{(s)}} ({b_{ij}}^{(s)} \cdot x_j^{(s)}\left(t_\lambda^{(s)}\right))$$

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.

3.6.3. DCM algorithm

Step 1: Preprocess the original data to obtain the stage-wise sample-based transcriptional data.

Step 2: Perform the gene-evolving trend analysis to determine the ascending or descending trend of each gene for each stage.

Step 3: Conduct the bootstrapping procedure and then produce the model equations of the inter-stage dynamical GRN.

(i) Obtain a random group of bootstrapping samples for each gene at each stage.

(ii) Produce the model equations using different settings for the fraction factor (λ) for each bootstrap group

(iii) Iterate (i) and (ii) to obtain the model equations.

Step 4: Estimate the model's coefficients of the ODE equation, and hence reconstruct the intra-stage dynamical GRN

(i) Based on the model equations obtained in Step 3, perform a LASSO or elastic net regression to solve the model coefficients with different network sparsities.

(ii) Determine the model coefficients that have the most appropriate sparsity for different stages using cross-validation approach

(iii) Reconstruct the GRNs described for each stage s by the ODE equation according to ${b_{ij}}^{(s)}$, which is proportional to $\beta_{ij}^{(s)}$.

Step 5: Repeat Step 3 and Step 4 to calculate the confidence of the network connection.

All the steps listed above is done in R (version 3.3.3) using packages such as parallel, annotate, hgu133a.db, igraph, and most importantly, glmnet package that allows to fast implementation of

LASSO or elastic net regression in selecting appropriate penalty coefficient λ (which is different from the λ used to describe time section) and equation coefficients.

4. Results

4.1. Description of the merged datasets

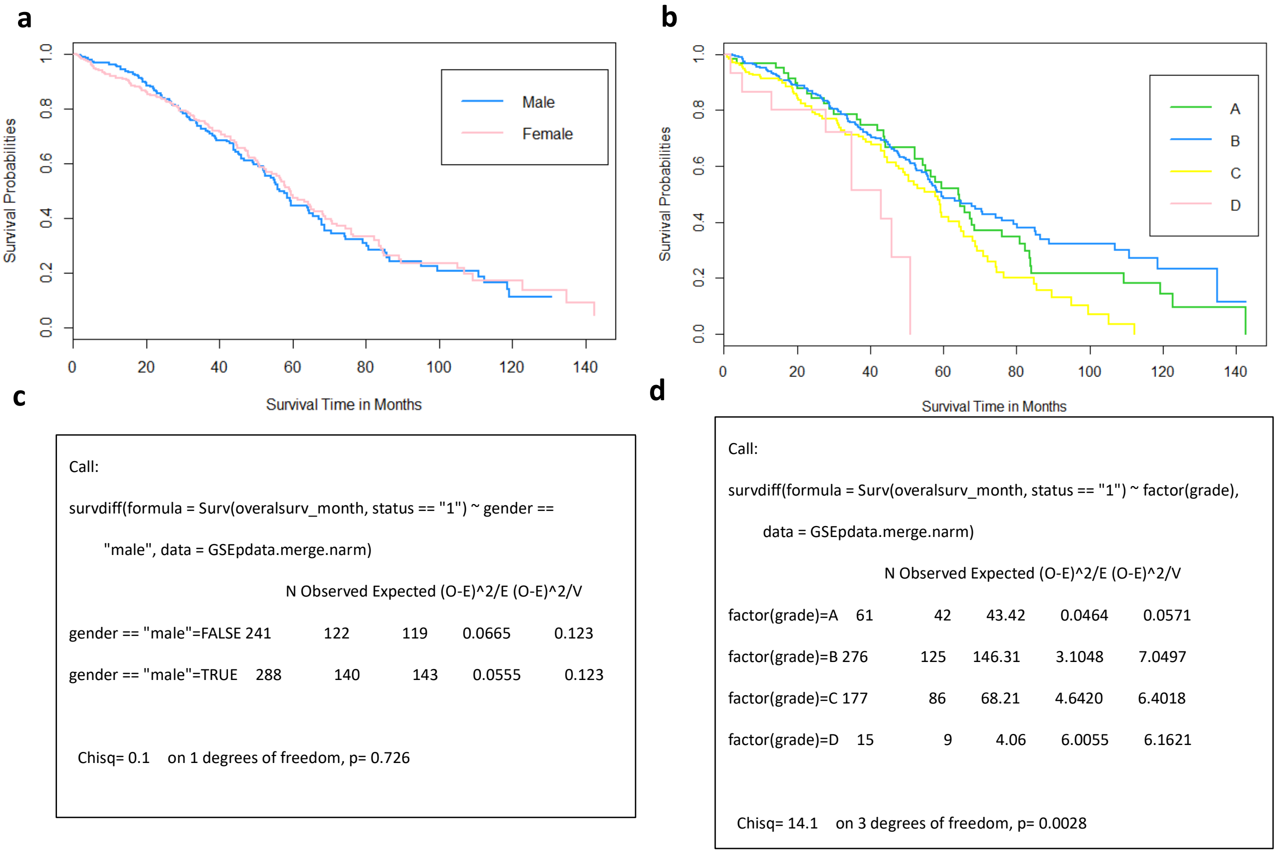
Because of the difference version of microarray chips, the GSE12945 dataset uses a lower version of the chip (with smaller Affymetrix id and smaller annotation file), while all the other 3 datasets (GSE17536, GSE14333, patient67) all share the same chip type. First is to merge the phenotype datasets. NAs are removed prior to merging, and each data entry (sample) will only be removed when NAs exist in one of the variables (overall survival time, status).

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

	67-Patient	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

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.

4.2. Kaplan-Meier survival curve



```
Call:

survdiff(formula = Surv(overalsurv_month, status == "1") ~ gender ==

    "male", data = GSEpdata.merge.narm)

                              N Observed Expected (O-E)^2/E (O-E)^2/V

gender == "male"=FALSE 241      122      119     0.0665     0.123

gender == "male"=TRUE   288      140      143     0.0555     0.123


 Chisq= 0.1   on 1 degrees of freedom, p= 0.726
```

```
Call:

survdiff(formula = Surv(overalsurv_month, status == "1") ~ factor(grade),

    data = GSEpdata.merge.narm)

                    N Observed Expected (O-E)^2/E (O-E)^2/V

factor(grade)=A   61       42    43.42    0.0464    0.0571

factor(grade)=B 276      125   146.31    3.1048    7.0497

factor(grade)=C 177       86    68.21    4.6420    6.4018

factor(grade)=D   15        9     4.06    6.0055    6.1621


 Chisq= 14.1   on 3 degrees of freedom, p= 0.0028
```

Figure 1. Result of Kaplan-Meier survival curve and log-rank test

In our merged datasets, K-M curve is used to demonstrate the time-dependent survival proportion trend classified by gender and grade respectively. **Figure 1a** shows the K-M survival curve classified by gender, and they are pretty close to each other. **Figure 1b** shows the K-M survival curve classified by cancer grades. Grade A and B share a similar survival curve, Grade C is a bit lower than Grade A and B in the long term, while Grade D drop drastically and ended before half of the range of the survival time. This result corresponds with the colorectal cancer background, which highly recommends early treatment of colorectal patients.

In order to compare different K-M curves, the log-rank test is used to calculate the chi-square (X^2) statistics for each event time for each group and sums the results. **Figure 1c** is the log-rank test done using the R survival package, in order to statistically compare the difference between 2 curves representing different gender in the merged dataset. P value for the test is 0.726 which means no statistical significance between 2 curves and accept the null hypothesis. **Figure 1d** is log rank summary for comparing 4 curves representing different grades. The p-value is 0.0028, which is much lower than 0.05, hence we tend to reject the null hypothesis and conclude that this curve is statistically different from each other.

4.3. Using Cox Proportional Hazard Model to select significant gene

CoxPH model described in the methods are used in this step to select the differentially expressed genes for downstream analysis. If a gene is believed to be significant in affecting the survival status of patients, then it is selected with a p-value less than 0.05. In this article, we use the CoxPH model with stratification of dataset source (from which we believe technical bias could be coming from) and select differentially expressed genes for different levels of the p-value. There are 4407 genes with p-value under 0.05; 1868 genes with p-value under 0.01; 1300 genes with p-value under 0.005; 528 genes with p-value under 0.001; 352 genes with p-value under 0.0005 and at last 123 genes with their p-values under 0.0001. Genes with a p-value under 0.0001 are selected for the appropriate number of genes for the downstream analysis and for less probability of Type I error.

4.4. Gene Ontology (GO) Enrichment Analysis

Gene Ontology project aims to standardize the representation of gene and gene product attributes across species and databases[10]. Gene ontology enrichment analysis is quite an ordinary analysis procedure in bioinformatics. 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/don't have the annotation and whether or not the genes are considered significant. We set up the p-value threshold to be 0.01 (under which there are 1868 significant genes, about 8.4% of total genes). topGO R package is used to do 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.

Table 2. GO Enrichment Analysis Result

GO.ID	Annotated	Significant	Expected	classicKS	elimKS	Term
GO:0051436	111	12	9.1	3.90E-05	3.90E-05	negative regulation of ubiquitin-protein...
GO:0051437	120	13	9.83	8.80E-05	8.80E-05	positive regulation of ubiquitin-protein...
GO:0051301	814	72	66.71	0.00011	0.00011	cell division
GO:0006273	13	4	1.07	0.00013	0.00013	lagging strand elongation
GO:0045740	104	18	8.52	0.00013	0.00013	positive regulation of DNA replication
GO:0030049	59	7	4.84	0.00013	0.00013	muscle filament sliding
GO:0035999	14	4	1.15	0.00014	0.00014	tetrahydrofolate interconversion
GO:0009157	14	3	1.15	0.00016	0.00016	deoxyribonucleoside monophosphate biosyn...
GO:0032596	12	3	0.98	0.00018	0.00018	protein transport into membrane raft
GO:0031145	111	12	9.1	3.90E-05	3.90E-05	negative regulation of ubiquitin-protein...

From **Table 2** we should see that the most enriched GO terms are closely related with the regulation of ubiquitin-protein, cell division, DNA elongation and replication and protein transportation, all of which is proven to be a critical process of cancer progression. To interpret the GO enrichment analysis result, we should say that the most significant genes have important gene function in those mentioned fields and that they draw our attention towards a more detailed picture of gene regulation in the cancer developmental process.

4.5. Protein-protein Interaction (PPI) with STRING

The STRING database[11] is one that provides protein-protein interaction analysis for various species, including direct (physical) and indirect (functional) associations. In addition to the modifiable interaction map shown in the website **(Figure 2)**, we can also extract as many network attributions such as node-to-node tsv files and high-resolution png plots. The PPI analysis with STRING is a next-step filter to select genes with known direct or indirect associations and narrow the range to select both statistically and biologically significant genes in colorectal cancer.

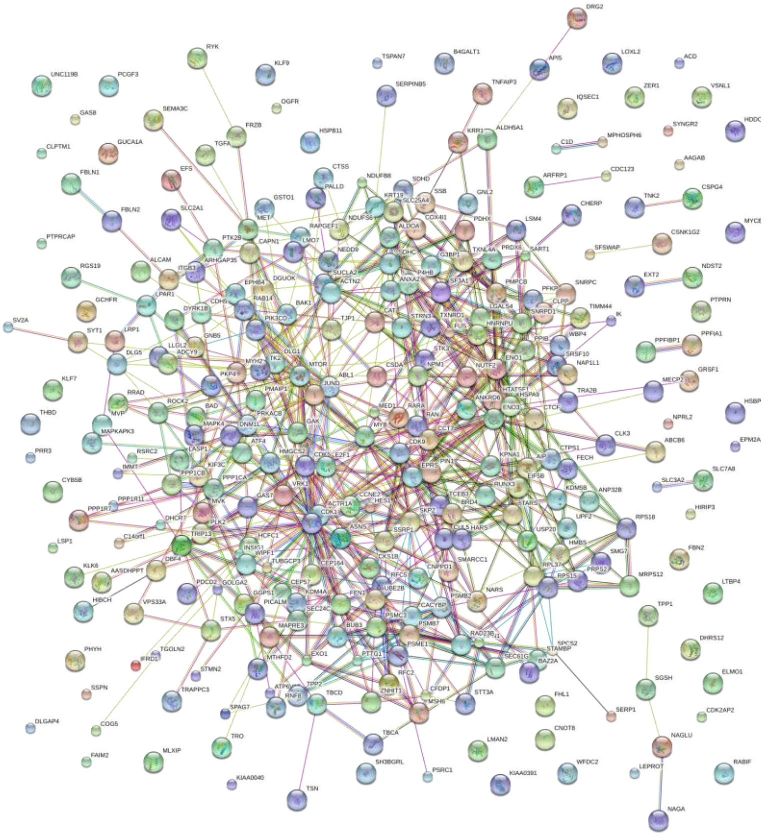


Figure 2. The PPI network visulization using STRING

Gene symbols (1868 of them) we have the previous CoxPH model, with a p-value under 0.01, are selected as the input to the STRING website. In a few minutes, the website returns with a huge PPI network plot with many connections as well as many single nodes (which need to be eliminated in the next step). The PPI link tsv file is downloaded and 249 unique genes are selected in consideration of potential biological meanings.

4.6. Heatmap Visualization

The heatmap in **Figure 3** is constructed using heatmap.3 source code from obigriffith in Github[12]. The function allows us to add the column bar to better visualize the connection between gene expression and patient information such as gender, cancer grade, and living status. The heatmap matrix value has been log 2 transformed, and from the density plot on the top left, the blue line shows the overall expression of the selected genes follows a normal distribution curve. Heatmap matrix is constructed from the 249 unique genes determined in the previous STRING PPI analysis. 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 heatmap is selected with 539 genes as rows and 529 patients as columns. From the heatmap we can see 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. However, on the left side of the heatmap, there are also living patients with much lower expression level comparing with the dead patients, who are in between. This feature needs further analysis to better understand the association between these genes.

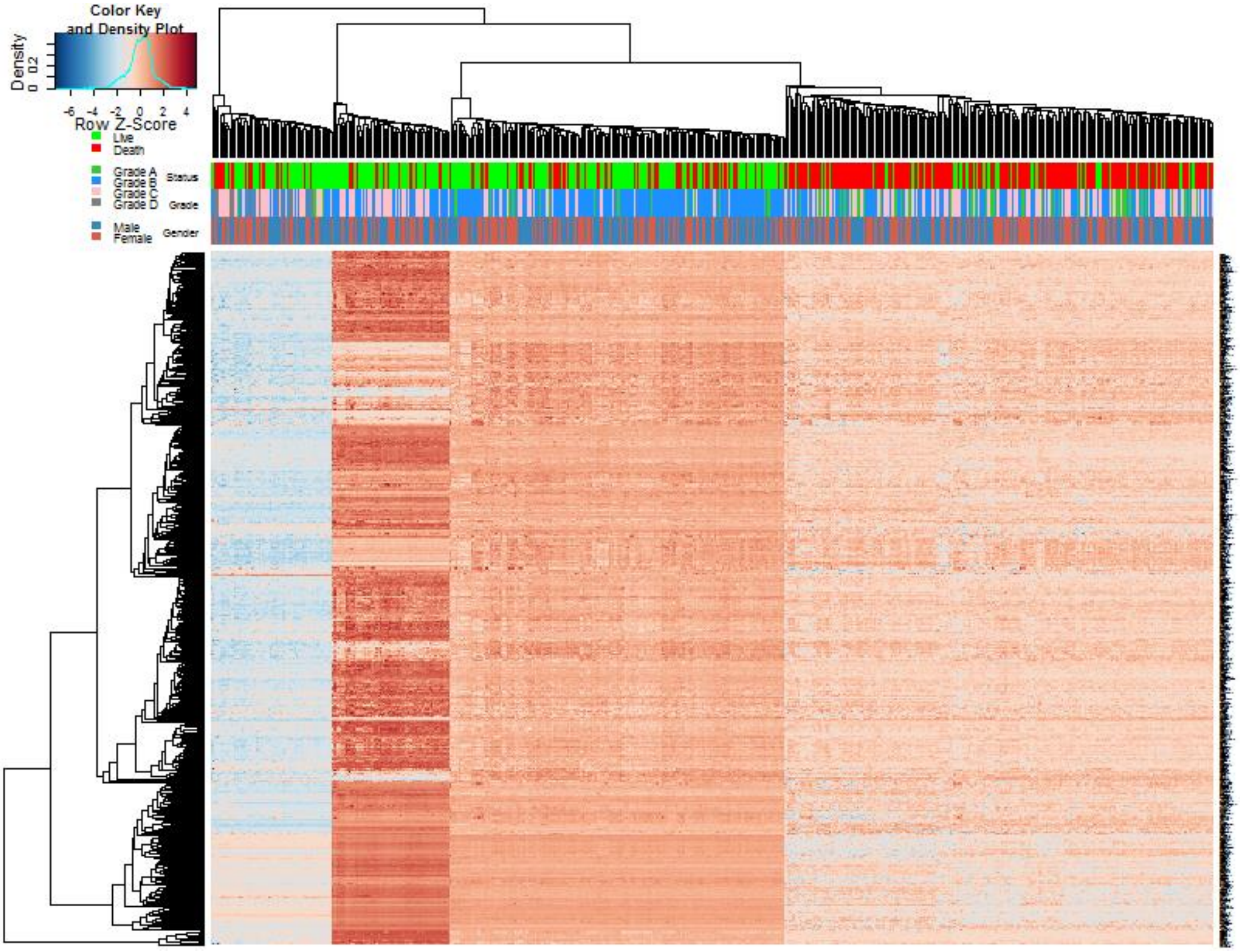


Figure 3. Heatmap plot of gene expression (539 genes of 529 patients)

4.7. Mutual Information Network

Using selected 249 genes from the previous STRING PPI network, mutual information network is built upon the minet R package, where we are able to use the ‘build.min’ function to generate the mutual information matrix. It computes the mutual information between all pair of variables according to the mutual information estimator. Then the ‘minet’ function is used to get the mutual information network with different inference algorithms such as CLR, ARACNE, and MRNET. After comparing different combinations of estimator and inference method (also known as sparsification method), we choose spearman estimator and ARACNE inference method for a detailed but not chaotic mutual information network. In order to demonstrate the network structure, fruchterman.regingold layout method is also used to automatically display the network. 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. Please note that in order to better visualize the connection, many duplicates with the same gene symbol are shrunk to a representative, which is chosen randomly.

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.

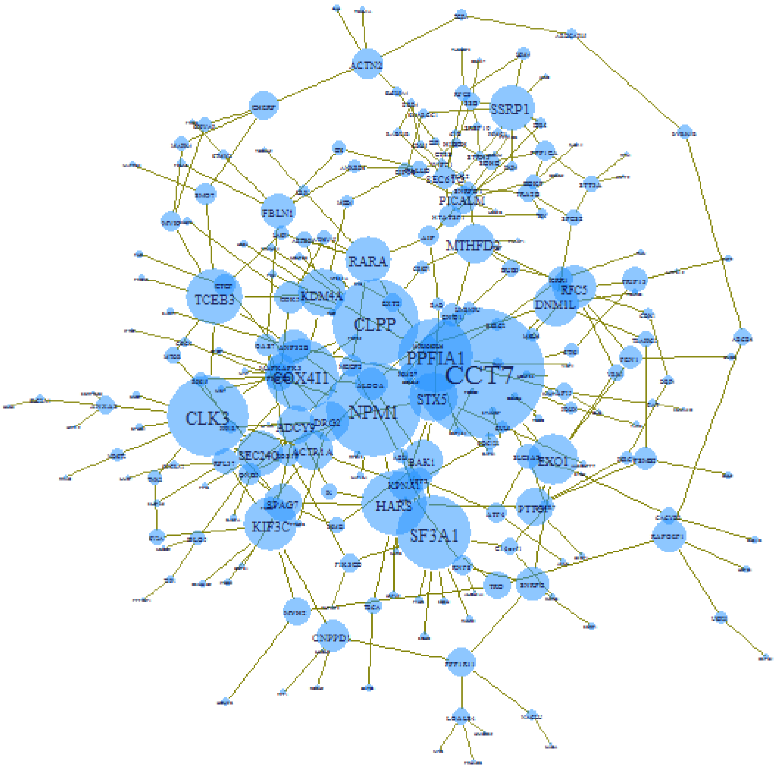


Figure 4. Mutual information network on 249 genes

From the mutual information network, we can tell gene CCT7 does not only has the largest text size but also it has the node size, meaning that CCT7 has both the largest connection with other genes and the sum of mutual information value. CCT7 is closely associated with endometrial carcinoma[13]. And now it has also been found to be most connected genes in CRC. The mutual information network enables us to have a big picture of the gene association within CRC, however, it cannot provide any information about the association direction or the level of such association.

Figure 5 shows the degree distribution **(Figure 5a)** and cumulative ones **(Figure 5b)** for the network. Most genes (about 45%) have only one connection while the most connected gene (NPEPL1) has up to 11 connections. **Table 3** shows the mutual information attributes for top 30 genes ranked by their p-values.

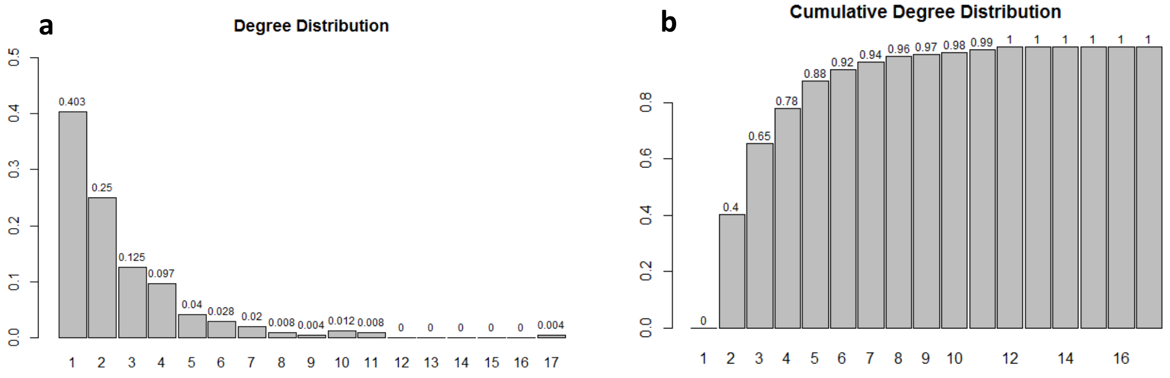


Figure 5. Separate and cumulative degree distribution of mutual information network

Table 3. Mutual Information Network Attributions for Top 10 Selected Genes (rank by degree)

Gene Symbol	Degree	Closeness	Betweenness	Eigen
CCT7	17	0.001534	6030	1
NPM1	11	0.001524	4702	0.982289
CLPP	11	0.001444	3154	0.642375
SF3A1	10	0.001263	3054	0.090174
PPFIA1	10	0.001445	3544	0.435155
CLK3	10	0.001297	3584	0.342707
COX4I1	9	0.001434	2710	0.650897
SSRP1	8	0.001165	749	0.000413
HARS	8	0.001365	2578	0.439195
MTHFD2	7	0.001465	5019	0.088204

4.8. Reconstruction the Gene Regulatory Network

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 p-value (<

0.01), PPI interaction and mutual information network (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.

Before the Dynamic Cascaded Modeling, a gene evolving trend analysis should be conducted in order to decide the gene expression evolving direction. 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:

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

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, $\mathrm{x}(t_1^{(s)})$ and $x\left(t_{N^{(s-1)}}^{(s-1)}\right)$ are the corresponding gene expression level of one gene. Accrding to the continuity assumption, $\mathrm{x}(t_1^{(s)})$ 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 6** illustrates the possible combinations between each stage.

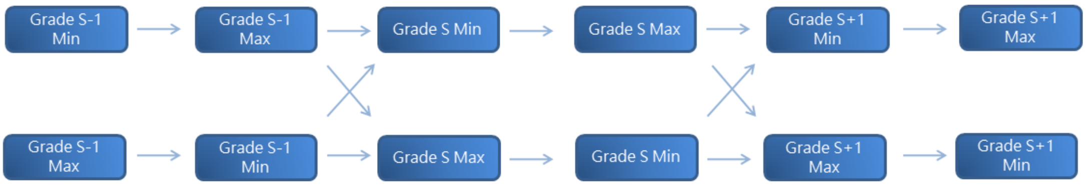


Figure 6. Scheme plot of possible gene evolving trends

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 sub lists 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 controling 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 are then estimated by LASSO or Elastic Net method. Cross validation strategy is also used to select the optimal λ (as the penalty coefficient). After getting the ${b_{ij}}^{(s)}$ matrix, we then convert them to $\beta_{ij}^{(s)}$ matrix according to previous formula. Then a gene regulatory network (GRN) is constructed according to these $\beta_{ij}^{(s)}$. The network contruction 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 the number of connections reserved in a network. The ‘glmnet’ uses a Elastic Net mixing parameter, with α_{ela} between 0 and 1. The penalty of the regression model is defined as

$$(1-\alpha_{ela})/2\|\beta\|_2^2+\alpha_{ela}\|\beta\|_1$$

When $\alpha_{ela}=1$, this mixing penalty is the same as LASSO penalty; while $\alpha_{ela}=0$ equals to ridge penalty.

A sequence set 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. The C-C plot with $\alpha_{ela}=0.5$ is selected to show the difference C-C curve of stages because it balance the ridge regression and the LASSO.

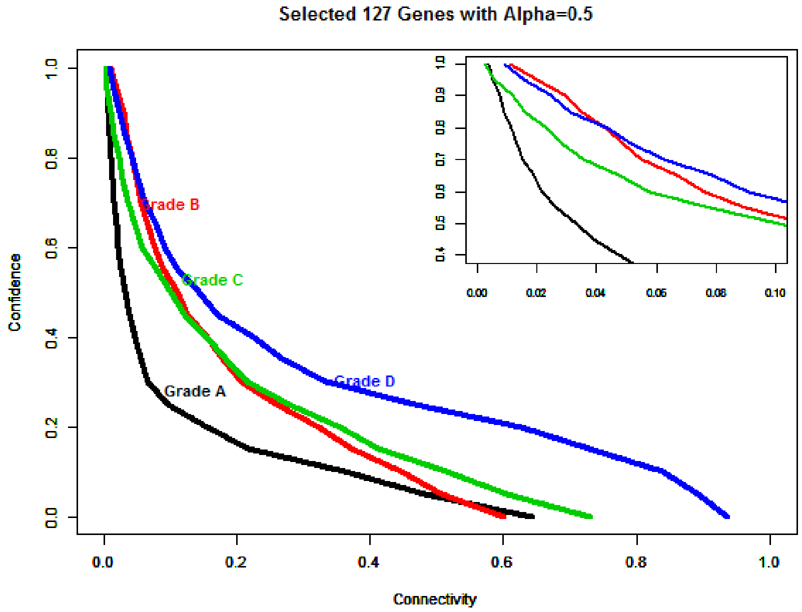


Figure 7. Connection confidence versus connectivity for four different CRC stages (grades)

From the **figure 7**, we notice that the GRN 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 127 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).

4.9. Network Enrichment Analysis

By implementing a threshold value of either network confidence or connectivity will help us build up the GRN, and the proper threshold will be determined using enrichment analysis, where a set of network connectivity is tested on whether the network connection is randomly selected or enriched for certain known interactions. These known interactions are extracted from 3 databases (the STRING, KEGG, and TRRUST database), then duplicates are excluded. Finally, 8794 known interactions are selected. Network enrichment was defined as the ratio of the proportion of known interactions over the baseline proportion of a random guess (in this case the baseline is 206 known interactions among all possible 16002 interactions). Network enrichment is considered significant if the two-sided proportional test (with the null model as the baseline proportion) has a pvalue under 0.05. Using different settings of network connectivity (PER), we get the network enrichment analysis in **Figure 8**. PERs are selected heuristically according to the Confidence-Connectivity curve. For a better network visualization purpose, we tend to set the network connectivity to no more than 0.05 (under which the network will have about 800 edges). Hence a set of PER from 0.01 to 0.05 is used for network enrichment analysis. From the plot below we can grab a rough impression about the enrichment status. For stage A and B networks, all PER settings have low connections enrichments which are under 50%; while stage C and D have average connections enrichment over 100%. In addition, only stage B and D networks have significant connections enrichment values. This result may be due to the limited number of known gene interactions in the current merged database, and the connection enrichment significance should not be treated as the golden standard in deciding the correctness or even the importance of the network.

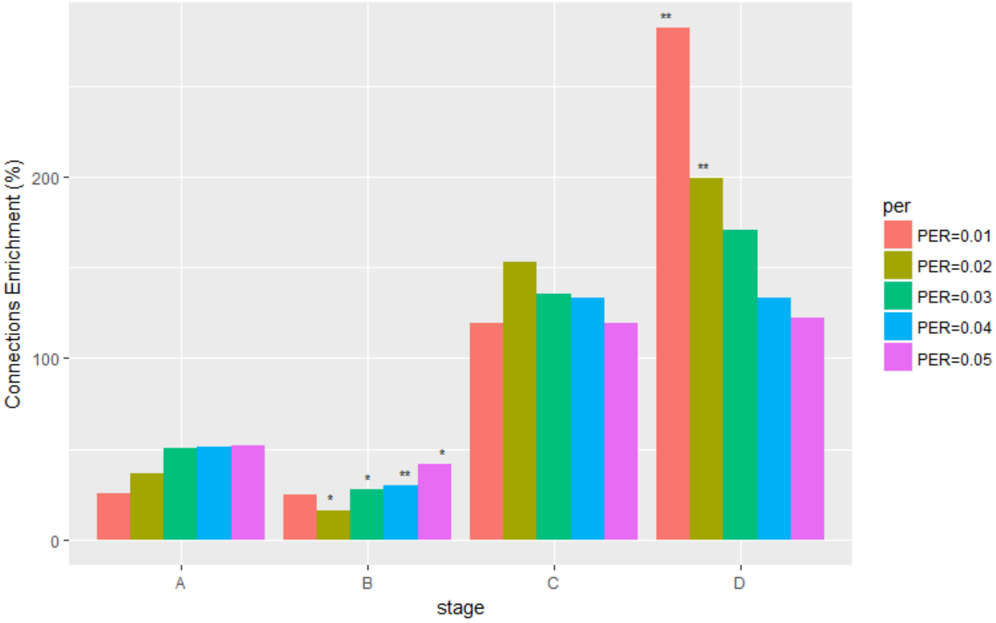


Figure 8. Network Enrichment Analysis

4.10. Functional Analysis

Four dynamic cascaded method-based (DCM) gene regulatory network (GRN) is reconstructed with the connectivity setting of 0.02. This setting is chosen out of the enrichment plot and of the consideration of network visualization. Four GRNs are then used by Cytoscape (http://www.cytoscape.org) and shown in **Figure 9**. 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 up regulation and down regulation 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. All known interactions are marked in green, while others in orange. The circular layout is ordered by the degree of the nodes.

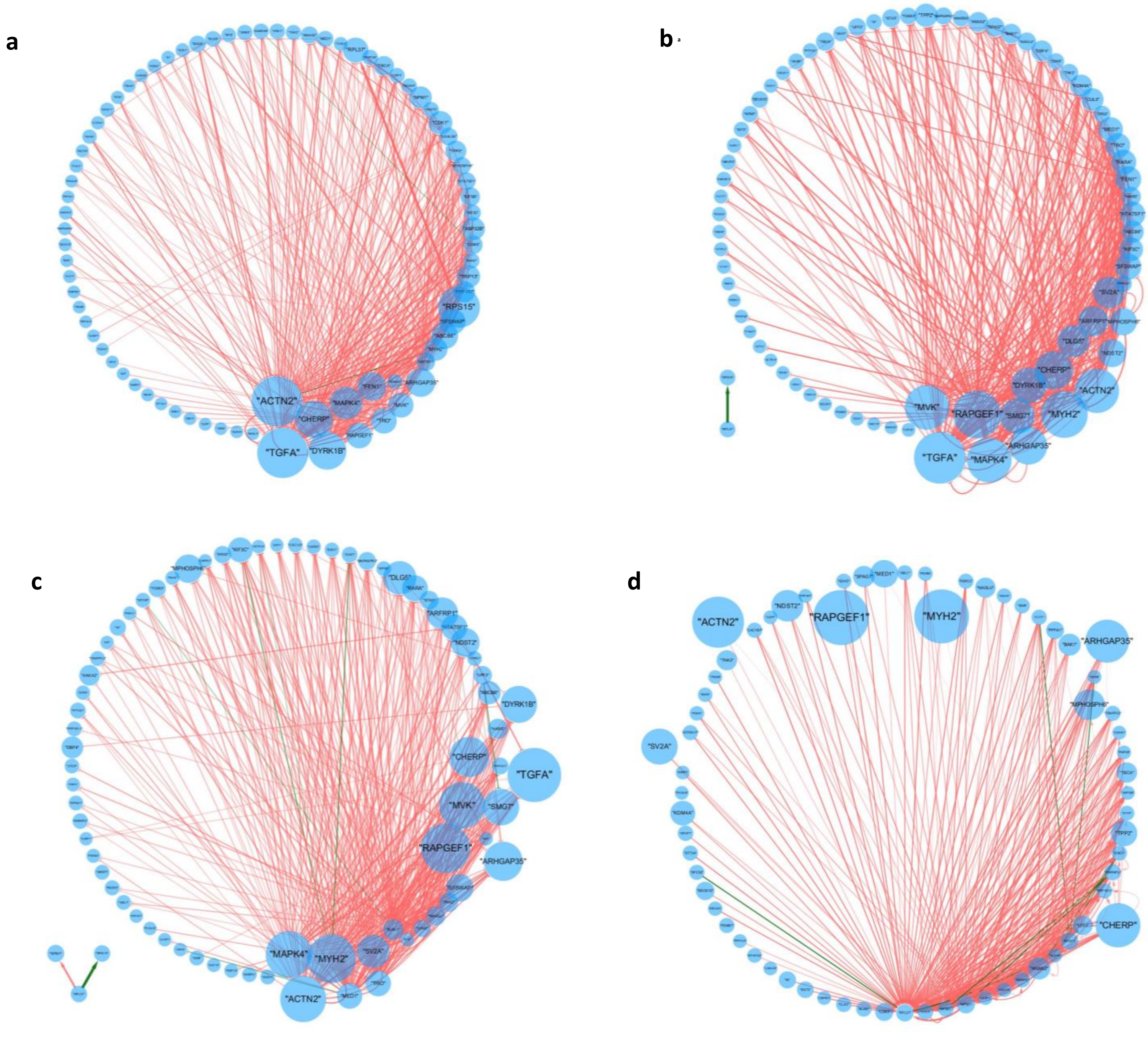


Figure 9. Reconstructing GRN based on four different CRC stages

The results show complex gene regulation behavior in all four CRC stages, and many connections are new and different from the known gene connections. Due to the difficulty of analyzing such complex regulatory network, a common filter is implemented on these four networks in order to unveil those genes and connections with the most significance.

The filter is set to select the edges (connections) with connection confidence larger than 95% (which means more than 95% of the parallel networks shares the same edge) and select nodes with sum of the products of the confidence and strength larger than 0.003. The sum of the products of the confidence and strength is associated with the gene activity in the network. Hence the filter intends to select a group of most active genes with high confidence level of connection with each other. **Figure 10** shows the four slimmer GRNs after the filter.

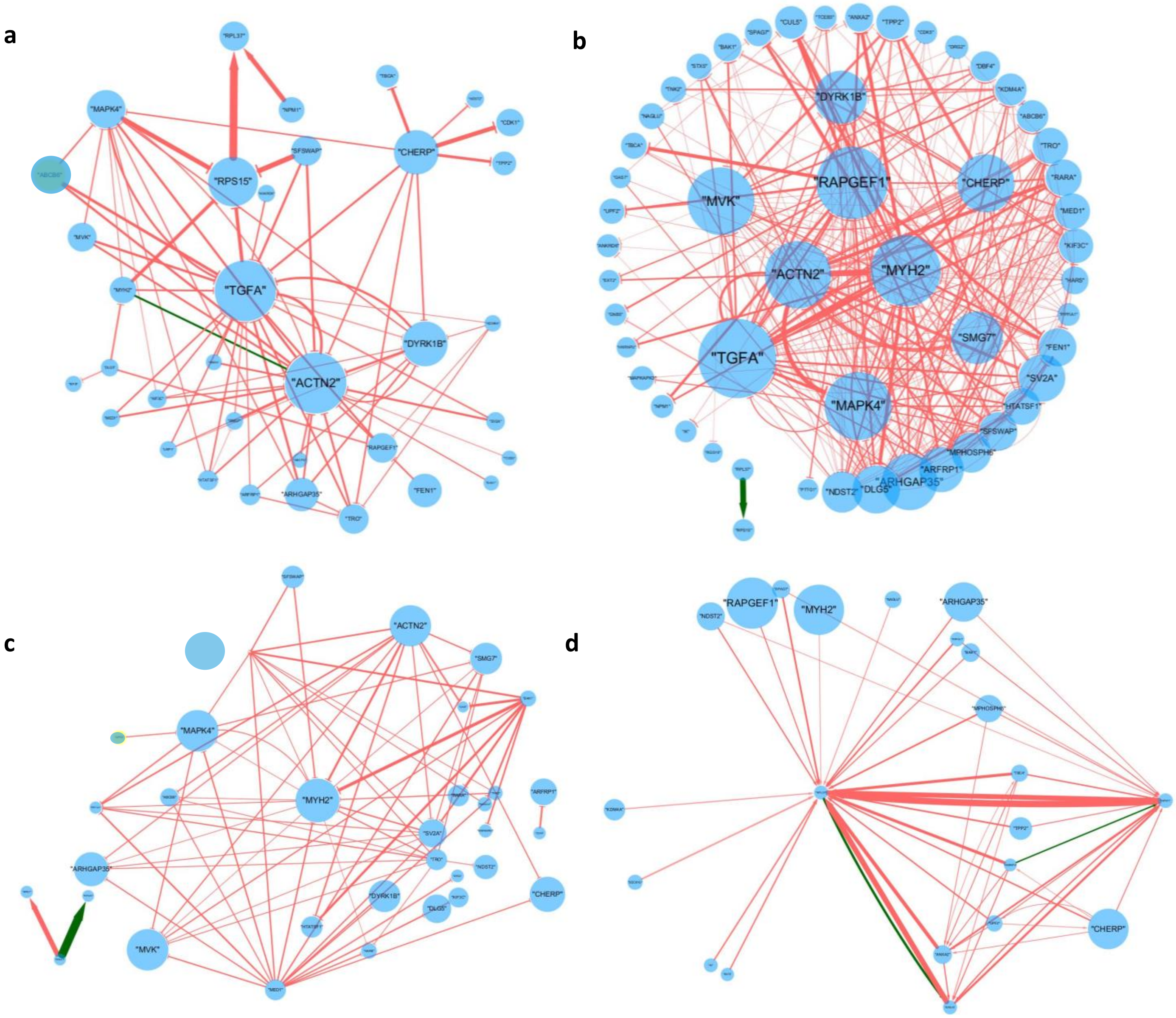


Figure 10. GRNs on four different CRC stages after filtering

After filtering, we are able to have a clear view of the most significant genes and the regulatory behavior between them. During CRC stage A **(Figure 10a)**, gene RPS15 has a big positive regulation over RPL37, and the former is down-regulating by SFSWAP, MAPK4 and other two genes. One known connection with high confidence level is the down-regulation from MYH2 to ACTN2. In CRC stage B, there are still many genes and many connections, so we put genes with the top connections within the circle. During stage B, 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[14]; ACTN2, as isoform of ACTN4, whose amplifications have worse outcomes than patients without[15], is also believe to be critical to the outcome of CRC patients； MYH2 is believed to be highly up-regulated (with 63 fold change) in CRC patients with tumors of high intrinsic COX-2 expression[16]; Research has shown the dependency of colorectal cancer on a TGF-beta driven program in stromal cells[17], and TGFA is of the same family of TGF-beta; CHERP is found in Colorectal Cancer Atlas (http://colonatlas.org/gene_summary?gene=CHERP) with function of negative regulation of cell proliferation. In CRC stage C, there is an interesting pattern of down-regulation from gene BAK1 to several other genes such as RAPGEF1, MYH2. Although BAK1 has relative 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 on the left-bottom of **Figure 10c**. This is reasonable because important regulatory genes are not always active while their effects could be tremendous. In **Figure 10d** showing the stage D regulation network, the previous phenomenon reoccurs as two less active genes RPL37 and NPM1 have big influence on the expression of each other as well as being regulated by various other active genes. From this pattern we hypothesize that those inactive but influential genes might be the downstream gene regulated by multiple less influential but more active genes. And the additive effect influence on the expression of the less active gene hence largely changes the downstream biological process.

Across all four stages, we notice a trend of network complexing from stage A to stage B, then sparsification from stage b to stage c, then to stage d. It’s possible that along the CRC progression, active genes are initialized and increased in the early step of cancer progression, then their number decreases as CRC progress to stage C and stage D, which might indicate the early decision process in CRC. When cancer is ready to change from early stage (stage A) to a middle stage (stage B), numerous genes are summoned and gene regulation network is also active and complex. When passing the middle stage, less active genes are needed, which might be the reason of passing the threshold of early cancer development towards a mature cancer progression stage. However, our model can only open the door for such a hypothesis and further experimental validation is needed to unveil the truth of gene regulation process in CRC.

5. Discussion

Various methods are used in this article to facilitate the selection and analysis of potential significant genes and dynamic gene regulation network built upon them. It’s been an excitement to reconstruct and visualize the dynamic gene regulation network with new DCM method. After all, our objective is to provide new prospective on CRC progression and help better understand such a horrible disease. The hypothesis theory provided based on the dynamic GRN is corresponding with the medical record of bad prognosis of middle or late CRC. In this paper, we also recommend unfortunate people to have regular physical examination and actively cooperate with medical practitioners to treat with early stage CRC. The early to middle CRC ‘threshold’ might be the last barricade to stop people from getting a lot worse outcome of such a disease.

Reference:

[1] Antonic V, Stojadinovic A, Kester K E, et al. Significance of Infectious Agents in Colorectal Cancer Development[J]. Journal of Cancer, 2012, 4(3):227-240.

[2] Staub E, Groene J, Heinze M, et al. An expression module of WIPF1-coexpressed genes identifies patients with favorable prognosis in three tumor types[J]. Journal of Molecular Medicine, 2009, 87(6):633-644.

[3] Jorissen R N, Gibbs P, Christie M, et al. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer.[J]. Clinical Cancer Research An Official Journal of the American Association for Cancer Research, 2011, 15(24):7642-7651.

[4] Smith J J, Deane N G, Fei W U, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer.[J]. Gastroenterology, 2009, 138(3):958-968.

[5] Goel M K, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate[J]. International Journal of Ayurveda Research, 2010, 1(4):274.

[6] Rich J T, Neely J G, Paniello R C, et al. A practical guide to understanding Kaplan-Meier curves.[J]. Otolaryngology Head & Neck Surgery, 2010, 143(3):331-336.

[7] 韩敏，梁志平. 改进型平均移位柱状图估算概率密度并对互信息作相关分析[J]. 控制理论与应用，2011, 28(6):845-850.

[8] Moon Y I, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators[J]. Physical Review E Statistical Physics Plasmas Fluids & Related Interdisciplinary Topics, 1995, 52(3):2318.

[9] Zhu H, Rao R S P, Zeng T, et al. Reconstructing dynamic gene regulatory networks from sample-based transcriptional data[J]. Nucleic Acids Research, 2012, 40(21):10657-67.

[10] Consortium T G O. Gene Ontology Consortium: going forward[J]. Nucleic Acids Research, 2015, 43(Database issue):1049-56.

[11] Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life[J]. Nucleic Acids Research, 2015, 43(D1):447-52.

[12] https://github.com/obigriffith/biostar-tutorials/blob/master/Heatmaps/heatmap.3.R

[13] Shan N, Zhou W, Zhang S, et al. Identification of HSPA8 as a candidate biomarker for endometrial carcinoma by using iTRAQ-based proteomic analysis[J]. Oncotargets & Therapy, 2016, 9(Issue 1):2169.

[14] Samuelsson J, Alonso S, Ruiz-Larroya T, et al. Frequent somatic demethylation of RAPGEF1/C3G intronic sequences in gastrointestinal and gynecological cancer[J]. International Journal of Oncology, 2011, 38(6):1575-7.

[15] Kazufumi H. The biological role of actinin-4 (ACTN4) in malignant phenotypes of cancer[J]. Cell & Bioscience, 2015, 5(1):41.

[16] Asting A G, Carén H, Andersson M, et al. COX-2 gene expression in colon cancer tissue related to regulating factors and promoter methylation status[J]. BMC Cancer, 2011, 11(1):238.

[17] Calon A, Espinet E, Palomoponce S, et al. Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation.[J]. Cancer Cell, 2012, 22(5):571-584.

Acknowledgement

It’s been such a long journey in my pursuing in Statistics. Thanks a lot to my tutor Dr. Xiaoqiang Sun, who has always been active online to ask my questions as well as ease my pain with comforts. Also thanks for all my friends who allow me to share my struggling feelings with them and keep me going on. Special thanks for Jianheng Liu who provided me with his Linux version R program. Last but not least, thank me myself because of my persistence and can-do attitude in thesis topic selection process. I really enjoy all the exciting moments in exploring and testing. I also value the pain I gain from every program failure, from every ugly plot, and from every figure that’s beyond my own interpretation. And my last thanks to all the readers who are so boring to even finish reading my cliché of acknowledgement.