دانشگاه تربیت مدرس

گروه علم داده‌ها

پروژه درس تحلیل شبکه‌های اجتماعی ۱۴۰۰-۱۳۹۹

موضوع:

داده‌کاوی در شبکه تنظیم رونویسی و کشف ژن‌های عامل سرطان ریه با رویکرد تحلیل شبکه

استاد راهنما:

دکتر بابک تیمورپور

پژوهشگر:

ساجده لشگری

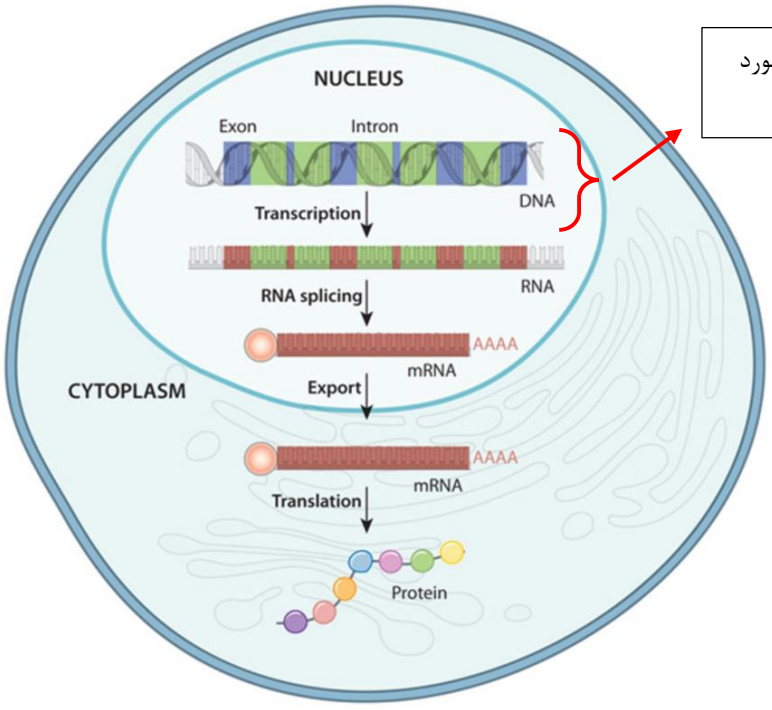
بهمن ماه ۹۹

فهرست مطالب

۱. چیستی مسئله و هدف از انجام پروژه

موضوع پروژه، کشف ژن‌های عامل سرطان ریه با رویکرد شبکه می‌باشد. برای اینکار از مرکزیت‌ها و اجتماع‌یابی استفاده می‌شود. هدف این پروژه، پیدا کردن ژن‌هایی است که دارای جهش‌های رونده در شبکه هستند که در اینجا ژن‌هایی را که در شبکه تنظیم رونویسی دارای مرکزیت درجه‌ای، میانداری و نزدیکی بیشتری هستند، به عنوان ژن عامل سرطان ریه (دارای جهشِ عامل سرطان) در نظر گرفته می‌شوند. همچنین به اجتماع‌یابی داده‌ها پرداخته و ژن‌هایی که دارای ارتباط بیشتری با هم هستند، از این طریق یافت می‌شوند.

شبکه تنظیم رونویسی، گونه‌ای از شبکه‌های زیستی است که از عوامل رونویسی و ژن‌های مختلف و برهم‌کنش آن‌ها ساخته شده است. در واقع عوامل رونویسی بر روی سایر ژن‌ها، اثر می‌گذارند. تحلیل این شبکه‌ها برای بررسی جریان اطلاعات در یک سامانه زیستی و شناخت مسیرهای مختلف که برای عملکردهای متفاوت، مفید است. گره‌ها در این شبکه، ژن‌ها و رونویسی‌ها هستند پس دو نوع پودمان در شبکه وجود دارد، پودمان ژنی و پودمان عامل رونویسی؛ و یال‌ها به معنی برهم‌کنش فیزیکی یا تنظیمی بین آن‌ها است. در پودمان نوع اول تعدادی ژن وجود دارد که همگی توسط یک عامل رونویسی تنظیم می‌شوند و در نوع دوم تعدادی عامل رونویسی هستند که همگی ژن‌های مشترکی را تنظیم می‌کنند.



شکل۱: شمای کلی سلول

در این پروژه برای یافتن ژن‌هایی که جهش آن‌ها موجب رخدادن سرطان می‌شود، از حداکثر میزان مرکزیت‌ها استفاده کرده، همچنین ژن‌های دارای ارتباط بیشتر با هم، با استفاده از سه الگوریتم پیدا شده اند که با توجه به میزان پودمانگی اجتماعات و الگوریتم بهتر شناسایی می‌شود و در بخش۵,۳ به طور کامل شرح داده می‌شود.

۲. مجموعه داده‌ها

ژن ناحیه خاصی از مولکولِ DNA با طول مشخص است. ژن‌ها که در هر سلولی یافت می‌شوند اطلاعات لازم برای تولید پروتئین‌ها را همراه خود دارند و با بیان این ژن‌ها، پروتئین‌های مختلفی تولید می‌شوند.

کنترل این فرآیندها در تعیین پروتئین‌های موجود در سلول و مقادیر آن، نقشی اساسی دارد. یعنی فرآیندی که سلول، رونویسی را روی RNA انجام داده تا مرحله‌ای که ترجمه روی mRNA انجام می‌شود و در نهایت پروتئین‌های تازه ساخته می‌شوند، بر میزان پروتئین بسیار تأثیر می‌گذارد.

شبکه تنظیم رونویسی (TRN[1])، شبکه اساسی برای کنترل فرایندهای سلولی است. تنظیم ژن، فعالیت ژن‌ها را در سطح رونویسی کنترل می‌کند. عوامل رونویسی (TF[2]) اجزای اصلی سلول هستند که مقررات را تنظیم می‌کنند. به عبارت دیگر، یک TRN نشان می‌دهد که چگونه هر یک از TFها، بیان سایر TFها و ژن‌ها را تنظیم می‌کنند. بسیاری از بیماری‌ها، از جمله سرطان از برخی اختلالات در عملکرد TFها ناشی می‌شود، که این مسئله اهمیت تجزیه و تحلیل TFها را در پژوهش‌های زیست پزشکی نشان می‌دهد.

در این پروژه از شبکه تنظیم رونویسی استفاده می‌شود و مجموعه داده‌های آن از پایگاه داده RegNetwork جمع‌آوری می‌شود که این شبکه مثالی از TRNهای وزندار شده، می‌باشد. در RegNetwork، لیست فعل و انفعالات نظارتی ژن از روش‌های مختلف و پایگاه داده‌های متعدد جمع‌آوری شده است. به طور خاص، در اینجا داده‌های انسانی TRN پایگاه داده RegNetwork مورد استفاده قرار می‌گیرد.

پس از انجام پیش پردازش‌های مطرح شده در بخش۳، مجموعه داده‌های بازیابی شده شامل 150202 فعل و انفعالات نظارتی و عوامل رونویسی (TF) و ژن‌ها (gene) است که همان گره‌های شبکه می‌باشند و روابط بین آن‌ها به صورت gene - TF و TF-TF است که هر کدام از این روابط بر اساس داده‌های بیان ژن سرطان ریه وزن‌دهی شده است و مقادیر وزن‌دهی با توجه به ستون CONFIDENCE در سایت http://www.regnetworkweb.org/search.jsp به ترتیب برای ۳ مقدار "کم"، "متوسط" و "زیاد"، ۰,۳، ۰,۵ و ۰,۸ در نظر گرفته شده است.

[1] Transcriptional Regulatory Network
[2] Transcription Factor

۱۰ داده اول در جدول۱ نمایش داده شده که ستون p نشان دهنده وزن یال‌های شبکه، Source گره مبدا و Target گره مقصد می‌باشد و روابط بین این گره‌ها، جهت‌دار در نظر گرفته شده پس گراف حاصل، گرافی جهت‌دار است و با انجام پیش پردازش‌های نهایی، گرافی با ۸۷۳۸۸ یال و ۱۱۰۱۶ گره و مقدار متوسط درجه داخلی و خارجی ۷,۹۳۲۸ به وجود می‌آید و این اطلاعات در جدول۲ نمایش داده شده است.

	Target	Source	p
0	ABL1	SHC3	0.8
1	ABL1	STAT5B	0.8
2	ABL1	CBLB	0.8
3	ABL1	CBLC	0.8
4	ABL1	CD55	0.8
5	ABL1	CRK	0.8
6	ABL1	CRKL	0.8
7	ABL1	RAC3	0.8
8	ABL1	RB1	0.8
9	ABL1	SHC1	0.8

جدول۱: شمای کلی از داده‌های نهایی مورد استفاده

```
Type: DiGraph
Number of nodes: 11016
Number of edges: 87388
Average in degree:   7.9328
Average out degree:   7.9328
```

جدول۲: اطلاعات شبکه مورد نظر

۳. پیش پردازش داده‌ها

شبکه RegNetwork انسانی، از سایت http://www.regnetworkweb.org/download.jsp جمع‌آوری شده و شمای داده‌ها به صورت جدول۳ می‌باشد. ستون اول و سوم به ترتیب مربوط به Target و Source می‌باشند، که ابتدا بقیه ستون‌ها را حذف کرده و این داده‌ها را به فرمت csv تبدیل کرده و لیست یال‌های اولیه در فایل 2links_data avalie قرار داده شده است. داده‌های بیان ژن سرطان ریه در فایل 3LUSC_nodes قرار دارند و شمای کلی این داده‌ها به صورت جدول۴ می‌باشد. همانطور که در جدول۴ در کادرهای قرمز رنگ مشخص شده، در برخی سلول‌ها چند نام گره به جای یک گره وجود دارد که با استفاده از تابع Exploding فایل Data manipulation.py، داده‌ها را دستکاری کرده و هر سلولِ فایل اکسل با استفاده از این تابع، مختص به یک گره می‌شود و مقداری که رو به روی آن، یعنی بعد از علامت "///" قرار دارد در سلول بعدی، زیر سلولِ فعلی، نوشته می‌شود و داده‌های فایل 4LUSC_nodes.modified حاصل می‌شوند. شمای کلی این داده‌ها در جدول۵ مشاهده می‌شود.

View - human.source

File Edit View Help

USF1	7391	S100A6	6277	
USF1	7391	DUSP1	1843	
USF1	7391	C4A	720	
USF1	7391	ABCA1	19	
TP53	7157	TP73	7161	
TP53	7157	SIAH1	6477	
TP53	7157	PMAIP1	5366	
TP53	7157	EI24	9538	
TP53	7157	CDKN1A		1026
TP53	7157	CCNG1	900	
TP53	7157	BAX	581	
TFAP2A	7020	VEGFA	7422	
TFAP2A	7020	TNPO1	3842	
TFAP2A	7020	JUP	3728	
STAT5A	6776	BCL2L1	598	
STAT3	6774	SOCS3	9021	
STAT1	6772	SOCS3	9021	
SRF	6722	FOS	2353	
SP1	6667	VEGFB	7423	

جدول۳: شمای کلی بخشی از داده‌های اولیه

	A	B	C
1	id	x1	x2
2	MIR4640///DDR1	9.185	10.529
3	RFC2	6.282	6.803
4	HSPA6	6.508	6.514
5	PAX8	8.945	8.898
6	GUCA1A	3.974	4.007
7	MIR5193///UBA7	8.272	7.508
8	THRA	5.734	5.851
9	PTPN21	5.328	4.861
10	CCL5	7.658	6.92
11	CYP2E1	4.059	4.133
12	EPHB3	7.362	8.17
13	ESRRA	7.799	7.723
14	CYP2A6	6.97	6.783
15	GAS6	10.951	10.202
16	MMP14	8.509	8.771
17	TRADD	8.01	8.153
18	CHURC1-FNTB///FNTB	6.223	6.618
19	PLD1	5.606	6.21

جدول۴: شمای کلی ۱۸ داده بیان ژن سرطان ریه

در مرحله بعد، داده‌های فایل 2links_data avalie بر اساس داده‌های موجود در فایل 4LUSC_nodes.modified فیلتر می‌شود و فقط گره‌های مشترک در هر دو فایل، باقی می‌مانند و داده‌های حاصل شده در جدول۵ نمایش داده شده اند.

علاوه بر TRNها، داده‌ها شامل فعل و انفعالات نظارتی microRNA (نمادهای شامل کلمات mir) نیز می‌باشد که در مطالعه حاضر حذف شده اند. پس از فیلتر کردن و رعایت این مورد، داده‌های جدول۱ حاصل می‌شود که در فایل 5WLUSC قرار داده شده و در ادامه، تحلیل‌ها روی این داده‌ها انجام می‌شود.

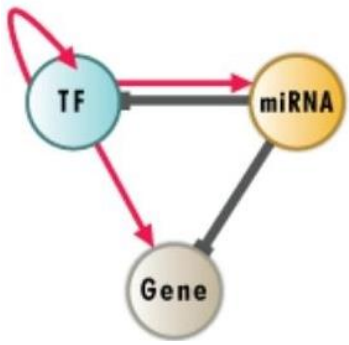
1	x1	x2	id
2	9.185	10.529	MIR4640
3	9.185	10.529	DDR1
4	6.282	6.803	RFC2
5	6.508	6.514	HSPA6
6	8.945	8.898	PAX8
7	3.974	4.007	GUCA1A
8	8.272	7.508	MIR5193
9	8.272	7.508	UBA7
10	5.734	5.851	THRA
11	5.328	4.861	PTPN21
12	7.658	6.92	CCL5
13	4.059	4.133	CYP2E1
14	7.362	8.17	EPHB3
15	7.799	7.723	ESRRA
16	6.97	6.783	CYP2A6
17	10.951	10.202	GAS6
18	8.509	8.771	MMP14
19	8.01	8.153	TRADD

جدول۵: شمای کلی ۱۸ داده بیان ژن سرطان ریه پس از دستکاری و فیلتر اولیه

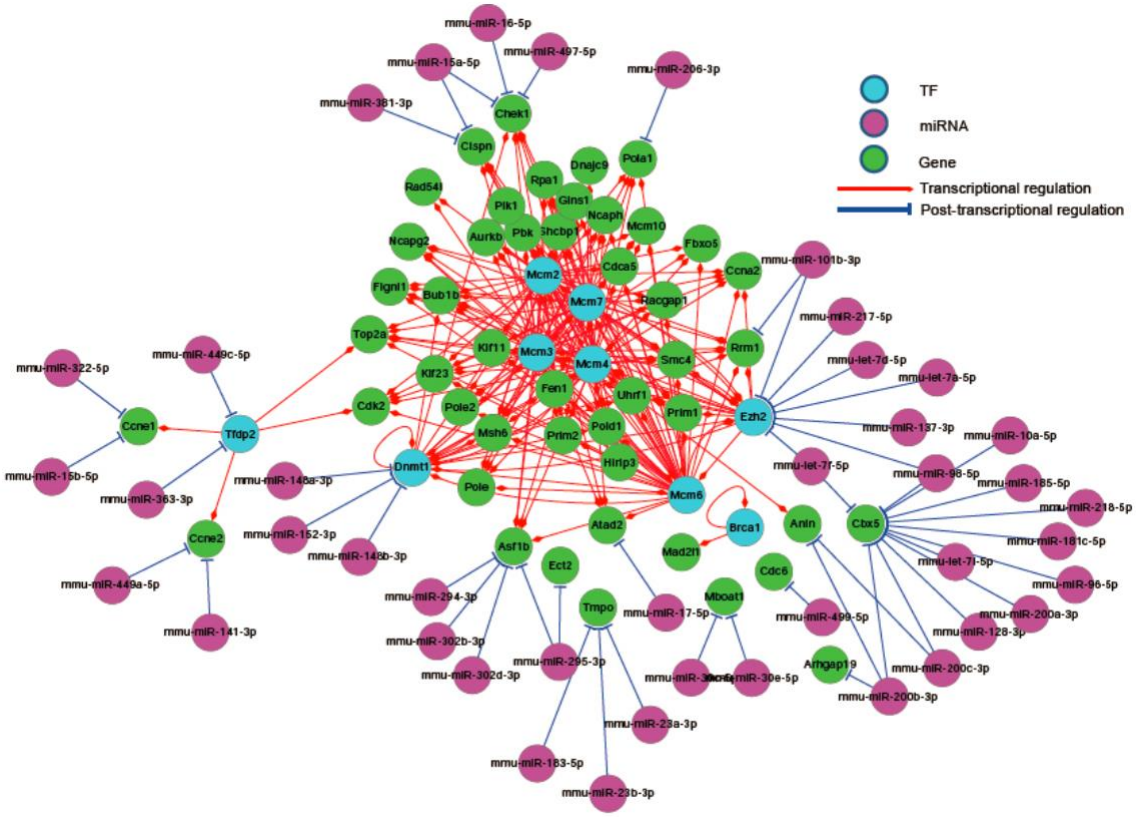
۴. مصورسازی شبکه

داده‌ها را با استفاده از برنامه Cytoscape و الگوریتم force – directed مصور کرده و شمایی از شبکه در شکل۲ قابل مشاهده می‌باشد.

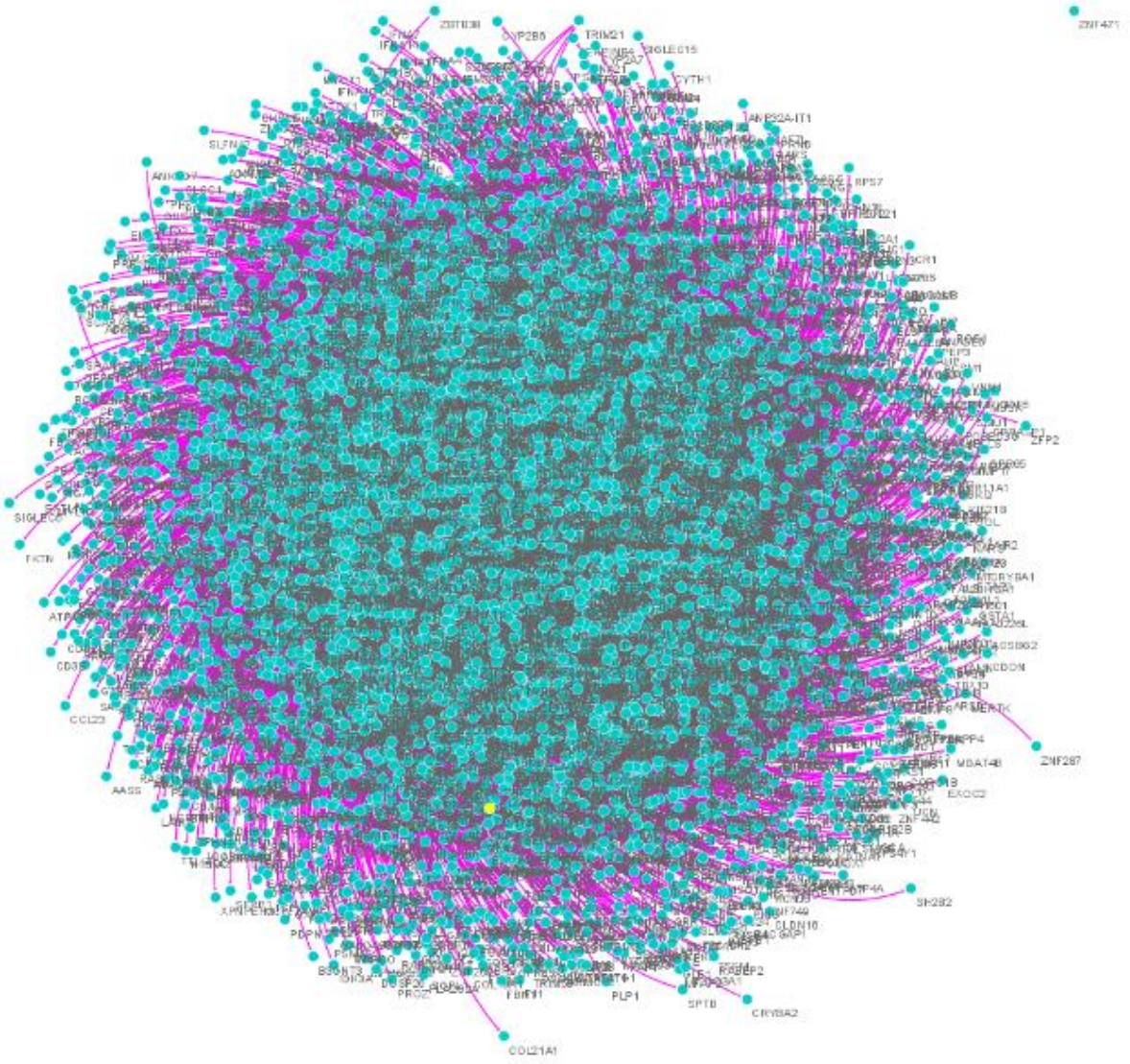
شکل کلی شبکه تنظیم رونویسی و پودمان‌های آن به صورت شکل۳ و ۴ است که در اینجا داده‌های مربوط به miRNA حذف شده اند.



شکل۳: شبکه تنظیم رونویسی در مقیاس بزرگ



شکل۴: شبکه تنظیم رونویسی در مقیاس کوچک‌تر



شکل۲: مصورسازی شبکه مورد نظر با استفاده از الگوریتم force – directed

گراف در نظر گرفته شده وزن‌دار، جهت‌دار و همانطور که در شکل۲ مشاهده می‌شود، ناهمبند می‌باشد و برای تحلیلِ شبکه چون در بیشتر مواقع شرط همبند بودن لازم است (مانند محاسبه مرکزیت‌های نزدیکی، میانداری و ...) ابتدا آن را به همبند تبدیل کرده و روی شبکه همبند تحلیل‌ها انجام می‌شود.

شبکه شامل ۲ مولفه همبند ضعیف و ۹۹۹۷ مولفه همبند قوی می‌باشد. برای همبند کردن شبکه از ماکسیمم مولفه همبند ضعیف استفاده می‌شود که اطلاعات این گراف در جدول۶ قابل مشاهده است.

باتوجه به تعداد گره‌ها (فقط یک گره، گره ZNF471، حذف شده) و نوع گره‌ها (گره یاد شده غیر عامل سرطان و بی‌اهمیت در این موضوع می‌باشد) شبکه حاصل برای تحلیل مناسب است.

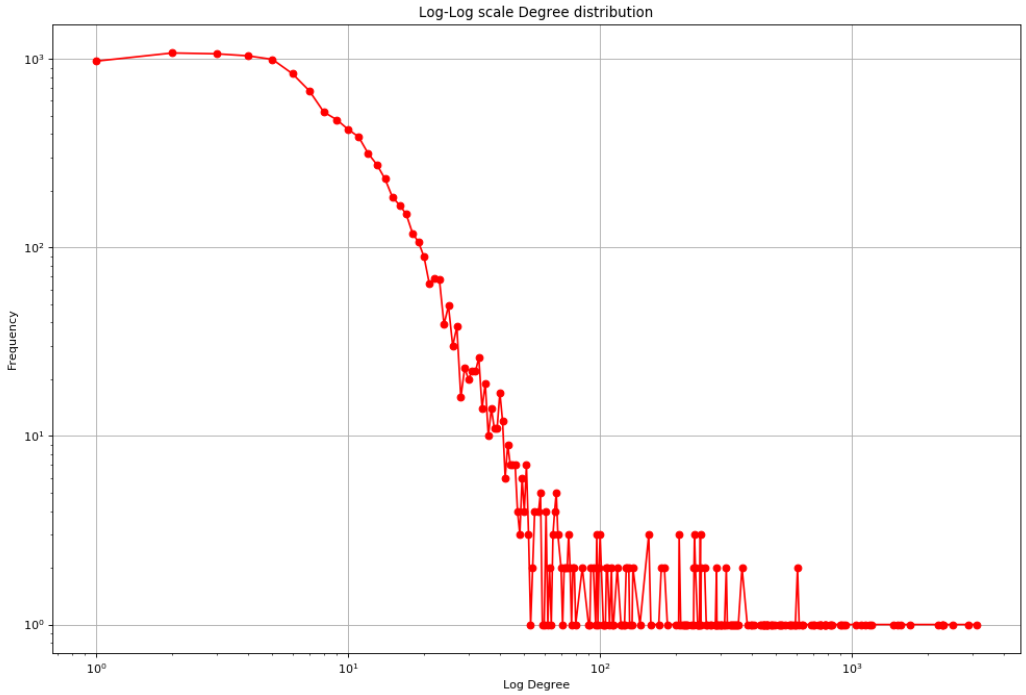
```
Type: DiGraph
Number of nodes: 11015
Number of edges: 87387
Average in degree:   7.9335
Average out degree:   7.9335
```

جدول۶: اطلاعات شبکه مورد استفاده در تحلیل

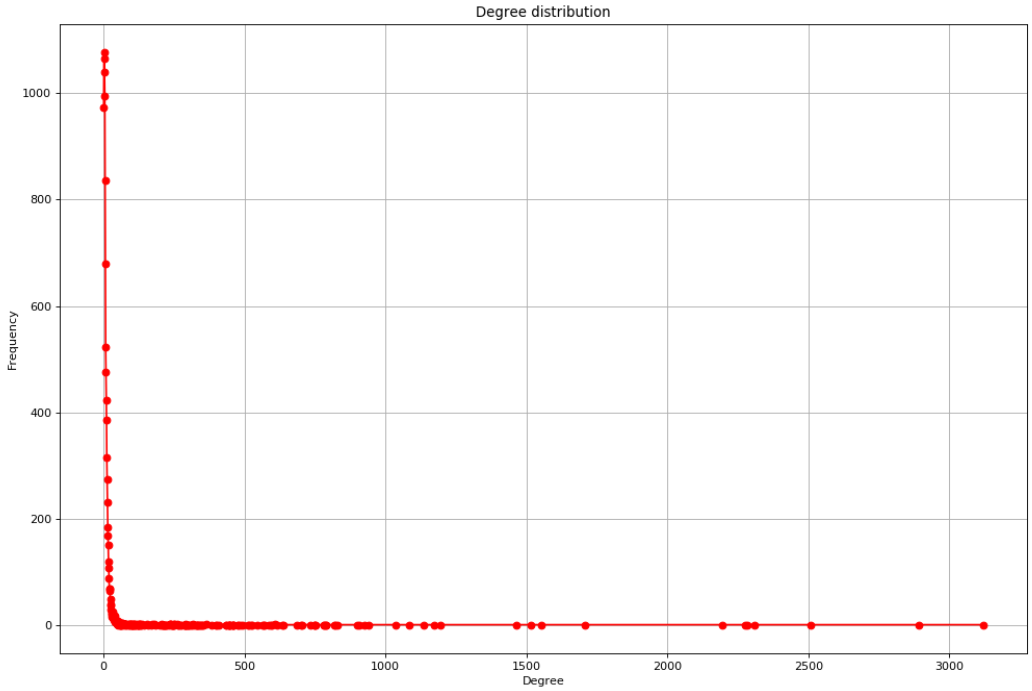
۵. تحلیل شبکه تنظیم رونویسی

۵،۱. بررسی توزیع و ساختار شبکه

توزیع درجات و مقیاس لگاریتمی آن، در شکل‌های ۵ و ۶ نمایش داده شده است. با توجه به شکل۵ و ۶، شکلِ ساختار شبکه به Scale Free شباهت بیشتری دارد؛ به ویژه در شکل۶. همچنین در شکل۵ نیز اگر گره‌های اول و آخر نادیده گرفته شوند، تابعی خطی مشاهده می‌شود که نشان می‌دهد توزیع شبکه به Scale Free شباهت دارد.



شکل۵: توزیع درجات شبکه در مقیاس لگاریتم- لگاریتم



شکل۶: توزیع درجات شبکه

جدول۷ نیز نشان دهنده این می‌باشد که شبکه موجود از حالت تصادفی دور است.

	برای شبکه مورد نظر	برای شبکه تصادفی
متوسط فاصله	۰,۲۹۷	۳,۳۶۷
متوسط ضریب خوشه‌پذیری	۰,۲۲۳	۰,۰۰۱۴

جدول۷: مقادیر متوسط ضریب خوشه‌پذیری و کوتاهترین فاصله

با توجه به اینکه توزیع شبکه Scale Free است با افزودن گره جدید به شبکه، انتظار می‌رود، گره جدید به گره‌ای که دارای بیشترین ارتباطات و درجه است، جذب و متصل شود.

همچنین گره‌هایی با درجه کمتر در شبکه، دارای فراوانی بیشتری هستند که به نظر می‌رسد، این گره‌ها دارای اهمیت بیشتری در شبکه باشند. از همین‌رو آن‌ها گزینه‌های شک برانگیزی برای عامل سرطان ریه بودن، در نظر گرفته می‌شوند که در ادامه با استفاده از معیارهای دیگر، بیشتر و دقیق‌تر مورد بررسی قرار می‌گیرند.

۵,۲. بررسی مرکزیت‌های شبکه

سه مرکزیت درجه‌ای، نزدیکی و میانداری برای این شبکه محاسبه شده و گره‌های مشترک بین حداقل دو مرکزیت و بیشتر، دارای اهمیت زیادی در شبکه می‌باشند زیرا اگر جهش عامل سرطانی در آن‌ها رخ دهند، به دلیل اینکه در مسیر ژن‌های بیشتری قرار دارند و یا دارای ارتباط بیشتری با گره‌های دیگر می‌باشند، همچنین دارای مسیر کوتاهتری با گره‌های دیگر هستند و رسیدن به این گره‌ها کم هزینه‌تر است، روی ژن‌های زیادی تاثیر می‌گذارند. از همین‌رو این گره‌ها (دارای مرکزیت‌های بیشتر) به عنوان ژن عامل سرطان ریه در نظر گرفته می‌شوند. در ادامه الگوریتم مورد استفاده به طور دقیق‌تر شرح داده می‌شود.

	مرکزیت درجه	مرکزیت میانداری	مرکزیت نزدیکی
حداقل مقدار	۰,۰۰۰۰۹	۰	۰
متوسط مقدار	۰,۰۰۱۴	$۱,۸۶ * ۱۰^{-5}$	۰,۰۲۹۵
حداکثر مقدار	۰,۲۸۳	۰,۰۰۹	۰,۵۴
گره مربوط به حداکثر مقدار	MAX	MYC	MYC

جدول۸: اطلاعات مربوط به مرکزیت‌های شبکه

در اینجا از دو الگوریتم برای یافتن ژن‌های عامل سرطان استفاده شده است.

الگوریتم اول:

۱. محاسبه مرکزیت درجه‌ای، نزدیکی و میانداری برای همه گره‌ها

۲. محاسبه میانگین برای هر سه مرکزیت

۳. تعریف مقدار گرد شده میانگین به عنوان آستانه یعنی به ترتیب برای مرکزیت‌های نام برده شده در خط ۱، ۰.۰۰۲، ۰,۰۰۰۰۱۹ و ۰,۰۳ (میزان گرد کردن متناسب با مقیاس هر کدام انجام شده)

۴. یافتن و جدا کردن داده‌های دارای مرکزیت بیشتر از آستانه به صورت جداگانه در هر مرکزیت

۵. یافتن و جدا کردن داده‌های مشترکِ موجود در هر سه مرکزیت‌ها

۶. فیلترگذاری روی داده‌های عامل سرطان واقعی و حذف داده‌هایی که در مجموعه داده‌های فعلی مورد استفاده وجود ندارند

۷. برچسب‌گذاری داده‌های واقعی حاصل شده در خط۶ با عدد ۱

۸. برچسب‌گذاری داده‌های پیش‌بینی شده بیان شده در خط۵ با عدد ۱ در صورت وجود در خط۶ و ۰ در صورت عدم وجود در خط۶

۹. یکسان‌سازی تعداد داده‌های واقعی و پیش‌بینی شده با استفاده از عدد ۰

۱۰. مقایسه داده‌های واقعی و پیش‌بینی شده و ارزیابی مدل با استفاده از معیارهای دقت، فراخوانی، صحت و مقدار- F

* الگوریتم دوم دارای مراحل یکسانی با الگوریتم اول است به جز خط۵ که به جای در نظر گرفتن داده‌های مشترکِ هر سه مرکزیت‌ها، از داده‌هایی که فقط در دو مرکزیت نیز وجود دارند، استفاده می‌شود.

با توجه به جدول۸ و الگوریتم دوم، گره MYC گره مهمی به حساب می‌آید و اگر دچار جهش رونده‌ای شود، در شبکه تاثیرگذار است. پس این گره به عنوان ژن عامل سرطان ریه در نظر گرفته می‌شود. همچنین جدول۹، نشان دهنده گره‌هایی هستند که در هر مرکزیت با توجه به مقدار آستانه مورد نظر، دارای بیشترین مقدار در همان مرکزیت هستند.

با توجه به خط۵ الگوریتم اول گره‌هایی که در هر سه بخشِ جدول۹ وجود دارند و بین هر سه، مشترک هستند را به عنوان ژن عامل سرطان ریه، انتخاب کرده که ۱۰۱۵ گره با بیشترین مرکزیت نزدیکی، ۴۹۲ گره با بیشترین مرکزیت میانداری و ۷۷۲ گره با بیشترین مرکزیت درجه یافت شده اند و در بین آن‌ها، ۴۲۷ گره مشترک وجود دارد که آن‌ها به عنوان ژن عامل سرطان در نظر گرفته شده اند و در جدول۱۰ نمایش داده می‌شوند. سپس آن‌ها را با ژن‌های عامل سرطان واقعی که در مجموعه داده‌ها موجود هستند (با تعداد ۴۹۳ گره)، مقایسه کرده و ارزیابی نهایی این الگوریتم، در جدول‌های ۱۱ و ۱۲ قابل مشاهده می‌باشد.

	Name
1	Name
2	ABL1
3	STAT5B
4	RB1
5	STAT5A
6	MYC
7	APBB1
8	APC
9	CTNNB1
10	TBL1X
11	AR
12	CITED2
13	ELF4
14	ELK1
15	ESR1
16	ESR2
17	ETS1
18	FHL2
19	FOXA1
20	GATA3

Closeness_Centrality

	Name
1	Name
2	ABL1
3	STAT5B
4	RB1
5	STAT5A
6	MYC
7	APBB1
8	APC
9	CTNNB1
10	DVL2
11	TBL1X
12	AR
13	VEGFA
14	MAPK3
15	MAPK1
16	HSP90AA1
17	AKT1
18	ATP2A2
19	BCL2
20	CASP3

Degree_centrality

	Name
1	Name
2	ABL1
3	STAT5B
4	RB1
5	STAT5A
6	MYC
7	APBB1
8	APC
9	CTNNB1
10	AR
11	CITED2
12	ELF4
13	ELK1
14	ESR1
15	ESR2
16	ETS1
17	FHL2
18	FOXA1
19	GATA3
20	GTF2F1

Betweenness_Centrality

جدول۹: بخشی از گره‌های با بیشترین مرکزیت

	Name
1	ABL1
2	STAT5B
3	RB1
4	STAT5A
5	MYC
6	APBB1
7	APC
8	CTNNB1
9	AR
10	CITED2
11	ELK1
12	ESR1
13	ESR2
14	ETS1
15	FHL2
16	FOXA1
17	GATA3
18	GTF2F1
19	HES1
20	

جدول۱۰: نمونه‌ای از ۲۰ ژن‌های عامل سرطان یافته شده با توجه به بیشترین مقدار مرکزیت‌ها با الگوریتم اول

```
confussion matrix Alg1:
[[113   0]
[399  94]]

clasification report for Alg1:
              precision    recall  f1-score   support

           0       0.22      1.00      0.36       113
           1       1.00      0.19      0.32       493

    accuracy                           0.34       606
   macro avg       0.61      0.60      0.34       606
weighted avg       0.85      0.34      0.33       606
```

جدول۱۱: ماتریس در هم آمیختگی و مقادیر ارزیابی الگوریتم اول برای هر کلاس

```
Accuracy1: 0.3415841584158416
Accuracy2: 0.3564356435643564
F1 score1: 0.32027257240204426
F1 score2: 0.375
Recall1: 0.19066937119675456
Recall2: 0.23732251521298176
Precision1: 1.0
Precision2: 0.8931297709923665
```

جدول۱۲: مقادیر ارزیابی مربوط به الگوریتم اول و دوم

همچنین با استفاده از الگوریتم دوم، ۶۰۶ گره مشترک به وجود می‌آید که آن‌ها به عنوان ژن عامل سرطان در نظر گرفته شده اند و برچسب‌گذاری برای ۴۹۳ داده واقعی با عدد ۱ و برای ۱۱۳ داده باقی‌مانده‌ی غیر سرطانی با ۰ انجام شده است. ارزیابی این ژن‌های عامل سرطان پیش‌بینی شده در جدول‌های ۱۲ و ۱۳ قابل مشاهده می‌باشد.

```
confussion matrix Alg2:
[[ 99  14]
[376 117]]

clasification report for Alg2:
              precision    recall  f1-score   support

           0       0.21      0.88      0.34       113
           1       0.89      0.24      0.38       493

    accuracy                           0.36       606
   macro avg       0.55      0.56      0.36       606
weighted avg       0.77      0.36      0.37       606
```

جدول۱۳: ماتریس در هم آمیختگی و مقادیر ارزیابی الگوریتم دوم برای هر کلاس

با توجه به الگوریتم اول ۱۱۳ ژن عامل سرطان و ۹۴ ژن غیر عامل سرطان به درستی پیش‌بینی شدند و ۳۹۹ ژن عامل سرطان بودند اما با توجه به الگوریتم به اشتباه، ژن غیر عامل سرطان پیش‌بینی شده اند. همچنین هیچ داده غیر عامل سرطانی با استفاده از این الگوریتم، به اشتباه عامل سرطان پیش‌بینی نشده است.

از طرفی با توجه به الگوریتم دوم ۹۹ ژن عامل سرطان و ۱۱۷ ژن غیر عامل سرطان به درستی پیش‌بینی شدند و ۳۷۶ ژن عامل سرطان بودند اما با توجه به الگوریتم به اشتباه، ژن غیر عامل سرطان و ۱۴ داده غیر عامل سرطانی، به اشتباه عامل سرطان پیش‌بینی شده اند.

به علاوه با توجه به جدول۱۱ و ۱۳ می‌توان نتیجه گرفت که الگوریتم اول در تشخیص ژن‌های غیر عامل سرطان ریه و الگوریتم دوم در تشخیص ژن‌های عامل سرطان ریه بهتر عمل می‌کند.

به طور کلی میزان دقت، فراخوانی و F-مقدار الگوریتم‌ها نشان می‌دهد که الگوریتم دوم بهتر از الگوریتم اول کار می‌کند و برای ساختن برنامه‌ای برای کلاس‌بندی ژن‌ها بهتر است از الگوریتم دوم استفاده شود.

ژن‌های عامل سرطان ریه یافت شده در فایل driver_found قابل مشاهده می‌باشند.

۵,۳. اجتماع‌یابی شبکه

برای اجتماع‌یابی نیز از شبکه همبند شده استفاده می‌شود و تک مشاهده ZNF471 را به عنوان داده‌ای دور افتاده که مربوط به ژن غیر عامل سرطان است و در این مسئله اهمیتی ندارد را حذف کرده و همانطور که پیش‌تر بیان شد، شبکه در نظر گرفته شده شبکه‌ای جهت‌دار می‌باشد بنابراین برای اجتماع‌یابی باید شبکه از حالت جهت‌دار به بدون جهت تغییر کند. با استفاده از الگوریتم حریصانه، لووین و انتشار برچسب به اجتماع‌یابی شبکه بدون جهت پرداخته و نتایج حاصل به صورت جدول۱۴ است.

	الگوریتم حریصانه	الگوریتم لووین	الگوریتم انتشار برچسب
تعداد اجتماعات یافت شده	۷	۱۳	۳
مقدار ماجولاریتی	۰,۲۹۵	۰,۳۰۶	۴,۸۰۳e−۰۵

جدول۱۴: اطلاعات حاصل از اجتماع‌یابی

باتوجه به مقدار پودمانگی نمایش داده شده در جدول۱۴ دو الگوریتم حریصانه و لووین تقریباً به خوبی عمل می‌کند و اجتماعات حاصل دارای گره‌های چگال می‌باشند و از حالت تصادفی تا حد خوبی دور هستند.

با استفاده از این الگوریتم، ژن‌های با نفوذ (انتشار در آن‌ها رخ می‌دهد) در هر اجتماع شناسایی می‌شوند و در نهایت مجموع ژن‌های با نفوذِ حاصل از همه اجتماعات، به عنوان ژن‌های عامل سرطان ریه پیش‌بینی شده و با استفاده از معیارهای استفاده شده در بخش ۵,۲، مورد ارزیابی قرار داده می‌شود. برای انتخاب الگوریتمِ بهتر، در اینجا (لووین و حریصانه) از مقایسه سرعت و دقت استفاده می‌شود.

576	2560
1521	2294
661	1355
722	2429
2541	551
1682	1452
1576	374
554	
648	
279	
44	
115	
96	

جدول۱۵: به ترتیب از راست به چپ، نشان‌دهنده تعداد گره‌های داخل هر اجتماع در الگوریتم حریصانه و لووین (اجتماعات مرتب هستند یعنی اعداد به ترتیب از بالا به پایین مربوط به اجتماعات ۱، ۲، ۳ ، ۴ و ... می‌باشند.)

باتوجه به نوع داده‌ها، اگر بخواهیم گروهی که شامل ژن‌های عامل سرطان هستند را پیدا کنیم، بهتر است به جای اجتماع‌یابی به خوشه‌یابی بپردازیم زیرا هدف پیدا کردن گروه‌هایی با ویژگی‌های شبیه به هم است نه گروه‌های چگال (بین ژن‌های عامل سرطان لزوماً ارتباطات زیادی وجود ندارد)، و برای انجام خوشه‌یابی به ویژگی‌هایی از ژن‌ها نیاز می‌باشد که با استفاده از آن‌ها خوشه‌یابی انجام شود.

اما هدف اصلی از انجام اجتماع‌یابی در این پروژه، بیش از اینکه یافتن اجتماعی شامل ژن‌های عامل سرطان باشد، یافتن اجتماعاتی چگال و مناسب است که در الگوریتم حداکثر انتشار مورد استفاده قرار گیرند.