مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره دوم. ص ۲۱۳-۲۰۱، ۱۳۹۹

بهبود استنتاج شبکه‌های تنظیم بیان ژن با رویکرد تجمیع داده‌ها

عاطفه ناصری[۱]، سیدمحمدحسین هاشمی‌نژاد[۲]*، مهران شرقی[۳]

• دریافت مقاله: ۹۸/۷/۱ • پذیرش مقاله: ۹۹/۳/۱۱

مقدمه: عمده‌ترین موضوع بر سر راه آینده بیوانفورماتیک طراحی ابزارهایی جهت مشخص کردن عملکردها و تمامی محصولات ژن‌های یک سلول است. این امر نیاز به ادغام رشته‌های متفاوت بیولوژیکی و همچنین ابزارهای پیچیده ریاضی و آمار دارد. در این تحقیق نشان داده شد که می‌توان با استفاده از تکنیک‌های داده‌کاوی مدل‌هایی برای تشخیص سبک زندگی افراد از لحاظ پرخطر یا کم‌خطر بودن برای ابتلاء به سرطان روده بزرگ توسعه داد.

روش: در این بررسی گذشته‌نگر، مجموعه داده‌ای شامل ۸۴ فرد بیمار و ۲۲۵ فرد سالم، شامل ۲۵ خصیصه جمع‌آوری شد. این اطلاعات شامل بیمارانی است که تشخیص آن‌ها مربوط به سال‌های ۱۳۸۵ تا سه ماهه اول ۱۳۹۳ می‌باشد. از پرکاربردترین تکنیک‌ها در ادبیات انفورماتیک پزشکی شامل ماشین بردار پشتیبان، بیزین ساده، درخت تصمیم و نزدیک‌ترین همسایگی برای توسعه مدل‌ها استفاده شد.

نتایج: مدل‌های توسعه داده شده با کارایی قابل قبولی، قادر به تشخیص سبک زندگی افراد هستند. سنجه غیرتکنیکی توسعه داده شده به خوبی می‌تواند ارزش واقعی تک‌تک پیش‌بینی‌ها، چه درست و چه نادرست را با هزینه‌های واقعی مشخص کند و یک میزان واقعی از هزینه‌های صرفه‌جویی شده در نظام سلامت توسط هر مدل را نشان دهد. از میان مدل‌های توسعه داده شده تنها دو مدل توانست معیارهای تعیین شده جهت استفاده در دنیای واقعی را ارضا کند.

نتیجه‌گیری: مدل‌های توسعه داده شده نه تنها باید از لحاظ تکنیکی ارزیابی شوند، بلکه باید از لحاظ سنجه‌های مورد پذیرش برای حوزه پزشکی و همچنین قابلیت اجرا برای حل واقعی مسئله نیز بررسی گردند.

کلید واژه‌ها: شبکه تنظیم بیان ژن، استنتاج شبکه تنظیم بیان ژن، الگوریتم انتشار، ادغام داده‌ها

•ارجاع: ناصری عاطفه، هاشمی‌نژاد سیدمحمدحسین، شرقی مهران. بهبود استنتاج شبکه‌های تنظیم بیان ژن با رویکرد تجمیع داده‌ها. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۲): ۱۳-۲۰۱.

۱. کارشناسی ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه الزهرا (س)، تهران، ایران
۲. دکتری مهندسی کامپیوتر، استادیار، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه الزهرا (س)، تهران، ایران
۳. دکتری مهندسی کامپیوتر، استادیار، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه الزهرا (س)، تهران، ایران

*** نویسنده مسئول:** سید محمدحسین هاشمی‌نژاد
آدرس: تهران، ده ونک، دانشگاه الزهرا (س)، دانشکده فنی و مهندسی
•شماره تماس: ۸۸۰۴۴۰۵۱- ۰۲۱ داخلی ۲۱۷۰
• **Email:** SMH.Hasheminejad@Alzahra.ac.ir

مقدمه

در دو دهه گذشته پیشرفت چشمگیر علم در شاخه زیست‌شناسی مولکولی این امکان را فراهم آورده است تا محققان با صرف هزینه و زمان اندک داده‌های ارزشمندی از سلول‌های موجودات زنده به دست آورند. حجم عظیم این داده‌ها و گسترش روزافزون آن‌ها، نیاز به شاخه‌ای جدید در علم برای ذخیره، بازیابی و تحلیل مناسب آن‌ها را اجتناب‌ناپذیر کرده است. این دانش نوظهور که با نام بیوانفورماتیک شناخته می‌شود به عنوان یک دانش بین‌رشته‌ای تلاش می‌کند تا با استفاده از تکنیک‌های موجود در علوم کامپیوتر، ریاضیات، آمار و علوم مرتبط دیگر، مسائل مختلف زیست‌شناسی مولکولی را حل کند [۱].

عمده‌ترین موضوع بر سر راه آینده بیوانفورماتیک طراحی ابزارهایی جهت مشخص کردن عملکردها و میانکنش تمامی محصولات ژن‌های یک سلول است. این امر نیاز به ادغام رشته‌های متفاوت بیولوژیکی و همچنین ابزارهای پیچیده ریاضی و آمار دارد. برای فهم عمیق‌تر عملکردهای سلولی، مدل‌های ریاضی جهت شبیه‌سازی واکنش‌های بسیار متنوع داخل سلول و تعامل آن‌ها در سطح سلول مورد نیاز است. مدل‌سازی مولکولی از تمامی فرآیندهای سلولی را بیولوژی سیستم‌ها می‌نامند. رسیدن به این هدف، خیز مهمی را برای فهم کامل یک سیستم زنده ایجاد خواهد کرد. به همین خاطر است که شبیه‌سازی سیستم و یکپارچگی آن‌ها به عنوان آینده بیوانفورماتیک در نظر گرفته می‌شود. بدیهی است که مدل‌سازی چنین شبکه‌های پیچیده و پیش‌بینی در مورد رفتار آن‌ها چالش‌ها و فرصت‌های بزرگی را برای محققین بیوانفورماتیک ایجاد خواهد کرد [۱].

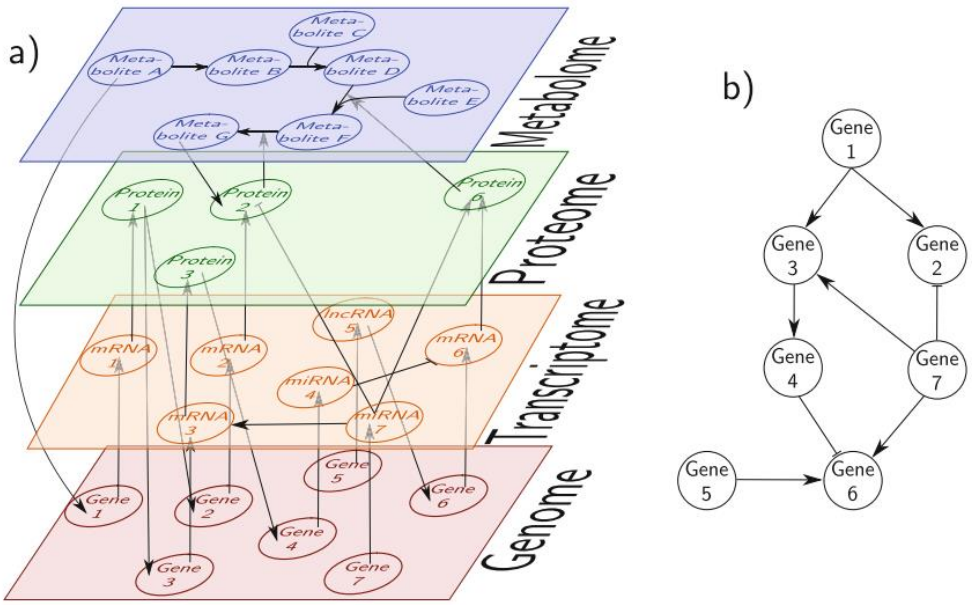
یکی از مهم‌ترین شبکه‌های پیچیده زیستی، شبکه‌های تنظیم بیان ژن هستند. بیان ژن یک فرآیند پیچیده از تبدیل یک توالی DNA به پروتئین مرتبط است که شامل رونویسی یک ژن و ساخت RNA به همراه رویدادهای پس ترجمه‌ای است که RNA به mRNA تغییر شکل داده و به مکانی می‌رسد که mRNA به زنجیره پروتئینی مربوطه ترجمه می‌گردد. تمام این مراحل تحت عنوان تنظیم بیان ژن شناخته می‌شوند. یک شبکه تنظیم بیان ژن (Gene Regulatory Network) GRN مجموعه‌ای از قطعات DNA در سلول است که با هم و با مواد دیگر در سلول ارتباط برقرار می‌کنند و در نتیجه میزان بیان ژن‌ها به mRNA در شبکه را کنترل می‌کنند [۲].

GRN با یک گراف نشان داده می‌شود. در گراف GRN هر رأس بیانگر یک ژن و وجود یال بین دو رأس مؤید مرتبط بودن دو ژن است. گره‌ها و یال‌ها با هم یک شبکه را تشکیل می‌دهند که اساساً توسط ساختار آن تعریف می‌شود، یعنی وجود یا عدم وجود ارتباط بین گره‌ها.

استنتاج شبکه‌های تنظیم بیان ژن شناسایی لینک‌های تنظیمی شبکه است؛ کشف این که کدام ژن‌ها به طور غیرمستقیم با یکدیگر ارتباط دارند (و چگونگی این تعاملات) با استفاده از میزان بیان mRNA و نهایتاً پروتئین‌های آن‌ها (شکل ۱). استنتاج شبکه تنظیم ژن‌ها یا یافتن روابط موجود بین ژن‌ها به وسیله داده‌های موجود یک مسئله پر اهمیت در زیست‌شناسی است، زیرا با کمک آن‌ها می‌توان به تشخیص بیماری‌های ژنی سهولت بخشید، بیومارکرهای مبتنی بر شبکه و داروهای جدید طراحی کرد و به مقایسه و بررسی الگوهای بیان ژن با عملکرد ناشناخته و به دست آوردن ایده‌هایی در مورد عملکرد آن‌ها پرداخت [۳].

همان‌طور که اشاره شد فرآیند تنظیم بیان ژن، یک فرآیند چند لایه‌ای است، از رونویسی تا سنتز پروتئین؛ اما اکثر مطالعات در حوزه استنتاج شبکه‌های ژنی تنها روی داده سطح mRNA (هنگام رونویسی) متمرکز شده‌اند و مکانیزم‌های رونویسی به عنوان داده اصلی برای محاسبات ژنتیکی سیستماتیک محسوب می‌شود. این در حالی است که در سال‌های اخیر این نتیجه حاصل شده است که بازسازی این شبکه‌ها تنها از روی یک نوع داده نمی‌تواند دقیق باشد و پذیرفته شده است که درک مفهومی سیستم بیولوژیک می‌تواند تنها با تحلیل ارتباط همه لایه‌های اُمیکس (Omics) به دست آید [۴]. چنین تحلیلی به ادغام داده اشاره دارد. ادغام داده، همه مجموعه داده‌ها را تحلیل کرده و یک مدل ارتباطی می‌سازد که همه مجموعه داده‌ها را هم‌زمان در بر می‌گیرد. نقطه شروع این تحلیل، استفاده از مفهوم ریاضی شبکه‌ها برای نمایش لایه‌های امیکس است. در دهه‌های اخیر، شبکه‌ها یکی از پرکاربردترین ابزارهای ریاضی برای مدل‌سازی و تحلیل داده امیکس بوده‌اند. خصوصاً این ابزارها به مطالعات شبکه ارتباطات پروتئین-پروتئین، شبکه‌های ارتباطات ژنی، شبکه‌های ارتباطی متابولیک و شبکه‌های بیان هم‌زمان ژن اعمال شده‌اند تا اطلاعات بیولوژیک ارزشمندی از ماشین مولکولی درون سلول استخراج شود. با این حال، درک کامل از سیستم بیولوژیک با تحلیل ادغام و ارتباط همه شبکه‌ها به دست می‌آید. ساخت یک شبکه یکپارچه برای گرفتن همه ارتباطات ژن-ژن با استفاده از

شبکه‌های مولکولی یکی از بزرگ‌ترین چالش‌ها در ادغام شبکه است [۵].



شکل ۱: شبکه تنظیم بیان ژن، نمایش بیولوژیکی در مقابل نمایش آماری از یک GRN

(a) سیستم‌های تنظیمی زیستی پیچیده هستند: محصولات واسط گوناگون ژن‌ها – رونوشت‌ها و پروتئین‌ها – و همچنین متابولیت‌ها در یک شبکه چندلایه تعامل دارند. چنین شبکه‌هایی بهترین نمایشی است که ما می‌توانیم از سیستم‌های پیچیده بیولوژیکی ارائه دهیم. (ب) نمایش آماری: ژن‌ها می‌توانند به عنوان گره‌هایی در یک گراف جهت‌دار در نظر گرفته شوند که در آن یال‌ها تعاملات تنظیمی را نشان می‌دهند. هر گره والد مستقیماً روی متغیرهای فرزند خود تأثیر می‌گذارد و مکانیسم تنظیمی یک محصول ژن را بر رونویسی ژن دیگر نشان می‌دهد [۶].

الگوریتم‌های مختلفی برای ادغام داده‌ها در این حوزه پیشنهاد داده شده‌اند. از یک منظر می‌توان روش‌های ادغام داده را بر اساس زیرساختی که برای حل مسئله استفاده می‌کنند، به ۴ دسته مختلف تقسیم کرد:

۱. روش‌های مبتنی بر شبکه
۲. مبتنی بر بیزین
۳. مبتنی بر کرنل یا هسته
۴. مبتنی بر فاکتورگیری غیرمنفی ماتریس

روش‌های مبتنی بر شبکه ساده‌ترین و راحت‌ترین راه برای یکپارچه کردن انواع مختلف داده در یک نمایش یکنواخت است. SNF (Similarity Network Fusion) [۷] یک مثال از ساخت شبکه‌های ترکیبی با ادغام داده همگن است. SNF که داده بیان mRNA, miRNA و داده متیلاسیون DNA را ادغام می‌کند، در ابتدا یک شبکه برای هر نوع داده ایجاد می‌کند و سپس یک شبکه یکپارچه کلی از ادغام همه شبکه‌های همگن تولید می‌کند. بیشتر روش‌های مبتنی بر شبکه روی تجمیع مجموعه‌ای از داده‌های ناهمگن در یک شبکه مجتمع واحد متمرکز شده‌اند که معمولاً با ترکیب یال‌ها در شبکه‌های مختلف از طریق استنتاج بیزین یا میانگین وزنی تطبیقی به دست می‌آید [۸–۱۰]. یک محدودیت اساسی در چنین رویکردهایی، از دست دادن اطلاعات قابل‌توجهی است که با نگاشت مجموعه داده‌های مختلف روی یک شبکه واحد ایجاد می‌شود.

نوع دیگر روش‌ها، مبتنی بر شبکه‌های بیزین هستند که متعلق به خانواده مدل‌های گرافیکی احتمالی می‌باشند. در پژوهش Isci و همکاران [۱۱] یک شبکه تنظیمی علّی احتمالاتی با ترکیب داده‌های از نوع بیان ژن، محل اتصال عامل رونویسی (Transcription Factor Binding Site) TFBS، (Expression of quantitative trait loci) eQTL و داده PPI (Protein-Protein Interaction) با استفاده از یادگیری بیزین ساخته شده است. شبکه بیزین ساخته شده از چند داده، قدرت پیش‌بینی بالاتری نسبت به مدل بیزین به دست آمده از تنها یک نوع داده (داده بیان ژن) دارد؛ اما دارای پیچیدگی بالای محاسباتی برای مجموعه داده‌های

بزرگ می‌باشد. در پژوهشی دیگر یک روش BN بر اساس انتخاب متغیر بیزین برای استنتاج GRN پیشنهاد شده است که از TFBS و PPI استفاده می‌کند [۱۲]. این روش از مدل خطی براساس این ایده استفاده کرده که وقتی یک شبکه دارای داده‌های آشوب باشد، تغییر گسترده در سطح بیان ژن‌ها به سرعت در کل شبکه پخش می‌شود. Žitnik و Zupan [۱۳] FUSENET را به عنوان الگوریتم استنتاج GRN براساس فرموله‌سازی شبکه مارکوف توسعه داده‌اند که چند مجموعه داده ناهمگن توزیع شده غیرمشابه را یکپارچه می‌کند. ورودی به FUSENET یک مجموعه از داده‌ها است که هر داده یک مجموعه از پروفایل‌های بیان ژن دارد. این داده‌ها سپس با داده‌های CNV و SNP ادغام می‌شوند. در مطالعه دیگری [۱۴] که با استفاده از زمینه‌های تصادفی مارکوف انجام شده است و دانش بیولوژیکی قبلی و داده ناهمگن را برای ساخت مدل‌های پیش‌بینی شبکه با صحت بالا استفاده می‌کند. این مطالعه از دو مجموعه داده برای ارزیابی صحت پیش‌بینی الگوریتم پیشنهادی بهره برده است. یکی از مجموعه داده‌ها A.Thaliana است. برای ساخت مجدد یک GRN، داده‌ها از سه منبع مختلف با هم ترکیب می‌شوند (۱) ناحیه پروموتر 2000bp از 17610 ژن (۲) پیش‌بینی اتصال DNA با این توالی‌ها برای 120 TFs و (۳) مجموعه داده بیان ژن A.Thaliana از نمونه‌های RNA از ۸۳ بافت. این داده‌ها برای استنتاج شبکه هم بیانی تحت شرایط خاص به کار می‌رود. یک فیلترینگ مبتنی بر واریانس برای حذف ژن‌هایی اعمال می‌شود که نشان دهنده واریانس کم در کل بافت‌ها و مراحل توسعه هستند.

دسته‌بندی بعدی یعنی روش‌های مبتنی بر کرنل، متعلق به روش‌های یادگیری ماشین هستند و نشان دهنده یک چارچوب ریاضی هستند که نقاط داده (ژن‌ها، پروتئین‌ها، miRNA و غیره) را از فضای ورودی I به فضای ویژگی F با اعمال تابع کرنل نگاشت می‌کند و این تناظر با ماتریس کرنل نشان داده می‌شود. این روش‌ها اولین بار در پژوهش Lanckriet و همکاران [۱۵] معرفی شدند که در آن Support Vector Machine) SVM برای مسئله طبقه‌بندی آموزش داده می‌شود و پروتئین‌های غشاء را از پروتئین‌های ریبوزوم جدا می‌کند. پژوهشگران مجموعه داده بیولوژیکی ناهمگن PPI، توالی‌های آمینواسید و داده بیان ژن را با استفاده از توابع هسته مختلف ادغام کرده‌اند. یافته‌های آن‌ها نشان داد که عملکرد این دسته‌بند روی مجموعه داده‌های یکپارچه در مقایسه با داده‌های منفرد بهتر است. در مطالعه دیگر برای استنتاج شبکه با استفاده از ادغام داده [۱۶]، چهار مجموعه داده متفاوت داده بیان ژن، داده تعاملات پروتئین، داده موقعیت‌یابی پروتئین و داده پروفایل فیلوژنتیک به ماتریس‌های کرنل متفاوت تبدیل شده‌اند.

دسته‌بندی چهارم فاکتورگیری غیرمنفی ماتریس، یک روش یادگیری ماشین است که برای حل مسائل خوشه‌بندی و کاهش ابعاد به کار می‌رود. هدف یافتن دو ماتریس غیرمنفی و کم بُعد است که حاصل‌ضرب آن‌ها تخمین خوبی از ماتریس ورودی غیر منفی باشد. به عنوان مثال در پژوهشی از روش فاکتورگیری غیر منفی ماتریس برای پیش‌بینی ارتباطات جدید ژن–بیماری با استفاده از دانش قبلی از شبکه مشابهت دارویی و شبکه PPI استفاده شده است [۱۹–۱۷].

در این پژوهش یک روش مبتنی بر شبکه استفاده شد که از فرآیند انتشار برای پویش توپولوژی شبکه بهره می‌برد و یک بردار ویژگی با ابعاد کم و درعین‌حال حاوی اطلاعات مفید برای گره‌های شبکه محاسبه می‌کند. در مرحله بعد، از این بردارهای ویژگی برای استنتاج شبکه‌های تنظیم بیان ژن استفاده می‌شود.

روش

همان‌طور که اشاره شد، روش‌های مبتنی بر شبکه بیشتر با ترکیب یال‌ها از شبکه‌های مختلف و تولید یک شبکه مجتمع واحد جزء پرطرفدارترین و ساده‌ترین روش‌ها برای ادغام داده‌ها هستند؛ اما یک محدودیت اساسی در چنین رویکردهایی، از دست دادن اطلاعات با نگاشت مجموعه داده‌های مختلف روی یک شبکه واحد می‌باشد. یک رویکرد برای مقابله با این چالش، تجزیه‌وتحلیل جداگانه ساختار هر شبکه و سپس ترکیب بردارهای ویژگی حاصل است. با این وجود، این رویکرد، ابعاد فضای ویژگی را به شدت افزایش می‌دهد که باید کاهش پیدا کند.

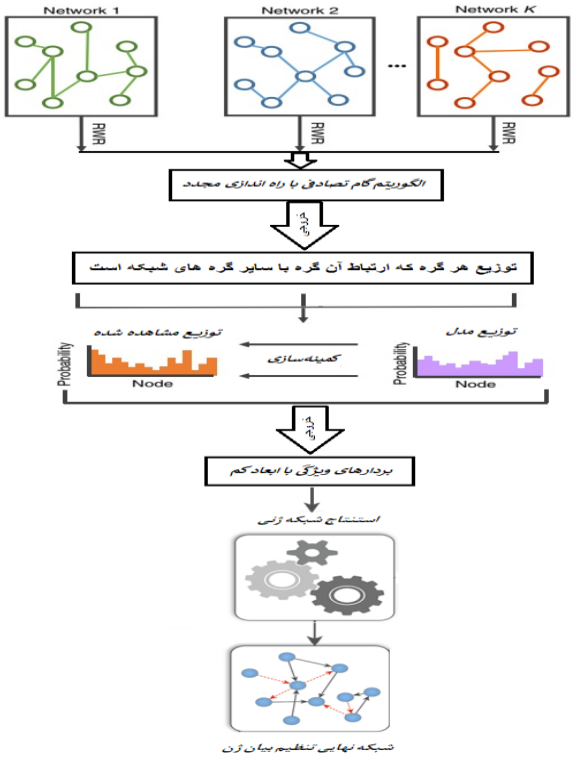
در این پژوهش، از یک چارچوب تحلیلی استفاده شد که ترکیبی از روش مبتنی بر انتشار (گام تصادفی با راه‌اندازی مجدد) و کاهش ابعاد برای استخراج بهتر اطلاعات توپولوژیکی شبکه به منظور تسهیل شناسایی عملکرد ژن‌ها است [۲۰]. ایده اصلی به دست آوردن ویژگی‌های حاوی اطلاعات مفید؛ اما با ابعاد کم است که خصوصیات توپولوژیکی ذاتی هر گره در شبکه را در بر می‌گیرد. الگوریتم‌هایی که قادر به پویش ساختار شبکه هستند، مثل انتشار شبکه یا گام تصادفی، نسبت به روش‌های پایه

عملکرد بهتری دارند [۲۱]. ایده اصلی این الگوریتم‌ها، انتشار اطلاعات در شبکه، به منظور بهره‌برداری از ارتباطات مستقیم و غیرمستقیم بین ژن‌ها است. این نوع از استراتژی بر این بینش تأکید دارد که ژن‌های متعامل و نزدیک به هم در توپولوژی شبکه، به احتمال زیاد عملکردهای مشابهی دارند.

می‌توان مراحل اصلی روش را به صورت زیر خلاصه کرد:

- ابتدا الگوریتم گام تصادفی با راه‌اندازی مجدد در هر شبکه اجرا می‌شود تا توزیع هر گره که ارتباط آن گره با سایر گره‌های شبکه است، به دست آید [۲۲].
- سپس هر یک از این توزیع‌ها با یک مدل لجستیک چند جمله‌ای برای هر گره تقریب زده می‌شود. با به حداقل رساندن واگرایی KL (Kullback-Leibler) (آنتروپی نسبی) [۲۳] بین توزیع‌های هر گره و توزیع لجستیک پارامتری چند جمله‌ای، بردارهای ویژگی با ابعاد کم برای هر گره به دست می‌آیند.
- بردارهای ویژگی تولیدشده به‌عنوان ویژگی‌های ورودی برای روش‌های استنتاج شبکه‌های تنظیمی ژن به‌کاربرده می‌شود تا به شبکه نهایی تنظیم بیان ژن دست یافته شود.

شکل ۲ چارچوب کلی فرآیند استنتاج ادغامی را نشان می‌دهد [۲۲].



شکل ۲: چارچوب کلی فرآیند استنتاج ادغامی

الگوریتم گام تصادفی با راه‌اندازی مجدد

الگوریتم گام تصادفی اولین بار برای پویش توپولوژی سراسری شبکه‌ها توسعه یافت که شامل شبیه‌سازی ذره‌ای است که به طور متناوب از یک گره به یک گره همسایه تصادفی منتقل می‌شود. اگر ذره را مقید کنیم که همیشه از یک گره یا مجموعه‌ای از گره‌ها مجدداً راه‌اندازی شود (هسته نامیده می‌شوند)، می‌توان از این الگوریتم برای اندازه‌گیری مجاورت بین هسته (ها) و تمام گره‌های دیگر شبکه استفاده کرد. به این طریق، یک توزیع از مشابهت توپولوژیکی برای هر ژن نسبت به سایر ژن‌های موجود در شبکه محاسبه می‌شود و از این راه می‌توان مرتبط‌ترین ژن‌ها را انتخاب کرد یعنی ژن‌هایی که مشابه‌ترین توزیع‌ها را دارند.

A ماتریس مجاورت یک شبکه تعاملی (وزنی) G=(E,V) با n گره است که هر گره یک ژن یا پروتئین را نشان می‌دهد. هر ورودی $B_{i,j}$ در ماتریس احتمال انتقال B که احتمال انتقال از گره i به گره j را ذخیره می‌کند، از طریق فرمول زیر محاسبه می‌شود:

$$B_{ij} = \frac{A_{ij}}{\sum_{i'} A_{i'j}}$$

رابطه (۱)

به‌طورکلی، RWR از گره i به این صورت محاسبه می‌شود:

$$s_i^{t+1} = (1 - p_r) B s_i^t + p_r e_i$$

رابطه (۲)

که در آن p_r احتمال راه‌اندازی مجدد است که تأثیر نسبی اطلاعات توپولوژیکی محلی و سراسری در انتشار را کنترل می‌کند. در حالت کلی e_i یک بردار توزیع n بعدی است به طوری که $e_i(i) = 1$ و $e_i(j) = 0,\ \forall j \neq i$ است. s_i^t یک بردار توزیع n بعدی است که در آن هر مؤلفه احتمال بازدید از یک گره را بعد از t مرحله نگه می‌دارد. عبارت اول در فرمول فوق مربوط به یک پال تصادفی است که به گره فعلی متصل است، در حالی که عبارت دوم مربوط به راه‌اندازی مجدد از گره اولیه i است. پس از چندین تکرار، s_i^t به یک توزیع پایدار همگرا می‌شود که این توزیع بیانگر احتمال انتقال از گره i به گره j است؛ بنابراین می‌توان نتیجه گرفت که اگر توزیع مشابهی بین گره i و گره j وجود داشته باشد، به این معنی است که آن‌ها موقعیت‌های مشابهی در شبکه دارند. از این رو، وقتی RWR همگرا می‌شود، بردار ویژگی حالت انتشار برای هر گره حاصل می‌شود.

کاهش ابعاد

بردار ویژگی حاصل از الگوریتم RWR، ابعاد بالایی دارد. با هدف کاهش نویز و کاهش ابعاد، هر حالت انتشار S_i با یک مدل لجستیک چندجمله‌ای تقریب زده می‌شود که ابعاد بسیار کمتری نسبت به حالت اصلی، n بعدی دارد. احتمال اختصاص داده شده به گره j در حالت انتشار گره i به صورت زیر محاسبه می‌شود:

$$\widehat{S_{ij}^k} := \frac{exp\{x_i^T w_j^k\}}{\sum_{j'} exp\{x_i^T w_{j'}^k\}}$$

رابطه (۳)

با توجه به این مدل، مسئله بهینه‌سازی به صورت زیر تنظیم می‌شود که شامل مجموعه‌ای از حالت‌های انتشار مشاهده شده $s = \{s_1, s_2, ..., s_n\}$ به عنوان ورودی می‌باشد و یک بردار ویژگی با ابعاد کم از گره‌ها، w و x که به بهترین وجه S را با توجه به مدل لجستیک چندجمله‌ای تقریب می‌زند، می‌یابد.

برای به دست آوردن w و x برای همه گره‌ها، از واگرایی Kullback-Leibler (آنتروپی نسبی) [۲۳] به عنوان تابع هدفی که باید مینیمم شود، استفاده شده است که یک انتخاب مناسب برای مقایسه توزیع‌های احتمال است [۲۰]:

رابطه (۴)

$$\text{minimize}_{w,x}\ C(s, \hat{s}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} D_{KL}\left(s_i^k \parallel \widehat{s_i^k}\right)$$

برای بهینه‌سازی تابع هدف، گرادیان‌ها با توجه به پارامترهای w و x محاسبه می‌شود:

رابطه (۵)

$$\nabla_{w_i^k} C(s, \hat{s}) = \frac{1}{n} \sum_{j=1}^{n} \left(\hat{s}_{ji}^k - s_{ji}^k\right) x_j$$

$$\nabla_{x_i} C(s, \hat{s}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{j-1}^{n} \left(\hat{s}_{ij}^k - s_{ij}^k\right) w_j^k$$

رابطه (۶)

روش شبه نیوتن استاندارد L-BFGS [۲۴] برای پیدا کردن بردارهای با ابعاد کم w و x متناظر با یک بهینه محلی استفاده شده است. مقداردهی اولیه بردارها بین اعداد تصادفی یکنواخت از [۰٫۰۵، ۰٫۰۵-] می‌باشد. نهایتاً در فاز آخر ویژگی‌های به دست آمده با ابعاد کمتر، به عنوان ورودی برای روش‌های مختلف استنتاج شبکه تنظیم بیان ژن استفاده می‌شود تا شبکه تنظیم بیان ژن نهایی ساخته شود.

تنظیم پارامترها

در آزمایش‌ها، الگوریتم‌های Regularized Gradient Boosting Machines (RGBM) و GENIMS (Gene Regulatory Network Inference Method based on a Multi-Level Strategy) با استفاده از بسته نرم‌افزاری R اجرا شده‌اند و الگوریتم‌های Genie3، TIGRESS و روش پیشنهادی در Matlab اجرا شده‌اند. تمامی پارامترهای قابل تنظیم الگوریتم‌ها، برابر با مقادیر پیش‌فرض مقاله‌ها قرار داده شده‌اند. در الگوریتم پیشنهادی مقدار احتمال راه‌اندازی مجدد در الگوریتم گام تصادفی برابر با ۰/۵ قرار داده شده است و تعداد ابعاد انتخاب‌شده برابر با ۳۰ است. تمام اندازه‌گیری‌ها با استفاده از Matlab نسخه ۲۰۱۷، پردازنده Intel Core (TM) i5 -3230U، ۲۶۰ گیگاهرتز، ۴/۰۰ گیگابایت حافظه رم و سیستم‌عامل ۶۴ بیتی مایکروسافت ویندوز ۷ انجام شد.

معیارهای ارزیابی

یکی از روش‌های مناسب برای ارزیابی نتایج حاصل از یک طبقه‌بند و ارزیابی میزان قابلیت آن در شناسایی طبقه موردنظر استفاده از منحنی Receiver Operating Characteristic (ROC) به‌منظور بررسی حساسیت روش است. برای رسم این منحنی، باید محور x که نمایانگر اختصاصی بودن (رابطه ۷) و محور y که نمایانگر حساسیت (رابطه ۸) است به ازای هر مقدار از آستانه طبقه موردنظر محاسبه گردد.

$$\text{اختصاصی بودن} = \frac{TP}{TP + FN}$$

رابطه (۷)

$$\text{حساسیت} = \frac{TN}{TN + FP}$$

رابطه (۸)

$$\text{دقت} = \frac{TP}{TP + FP}$$

رابطه (۹)

علاوه بر این، Area under the Precision-Recall curve (AUPR) که ناحیه زیر منحنی دقت-حساسیت را محاسبه می‌کند، نیز استفاده شده است.

نتایج

دو ارگانیسم واقعی Saccharomyces cerevisiae و باکتری Escherichia coli به عنوان مطالعه موردی انتخاب شده‌اند. برای هر مطالعه موردی مجموعه‌ای از شبکه‌های ژنی از بانک اطلاعاتی STRING [۲۵] استخراج شده است. هر مجموعه شامل هشت شبکه ناهمگن بر اساس انواع منابع داده از جمله ارتباطات همسایگی، اطلاعات ادغامی از پروتئین‌ها در گونه‌های دیگر، اطلاعات فلوژنیکی، اطلاعات هومولوژی، اطلاعات هم بیانی، داده‌های آزمایشگاهی تجربی، اطلاعات به دست آمده از پایگاه داده و اطلاعات حاصل از متن‌کاوی به

دست آمده‌اند. هر شبکه دارای به ترتیب ۳۳۳ و ۳۳۴ ژن است که تعداد لبه‌ها از ۱ تا ۱۱۰۰۰ متغیر است. ژن‌ها به فاکتورهای رونویسی شناخته‌شده (Transcription Factor)TF)) محدود شده است یعنی ژن‌هایی که به‌عنوان ژن‌های تنظیمی شناخته می‌شوند [۲۶] و تمام تعاملات تنظیمی بین این TFهای شناخته شده قلمداد می‌شود. مجموعه داده بیان ژن و همین‌طور شبکه استاندارد برای این ارگانیسم از کنسرسیوم DREAM5 [۲۷] استخراج می‌شود. برای هر مطالعه موردی این هشت شبکه به عنوان ورودی به الگوریتم پیشنهادی داده می‌شود و برای هر ژن یک بردار ویژگی محاسبه می‌شود. سپس این بردار ویژگی به یک الگوریتم استنتاج شبکه تنظیم بیان ژن داده شده و شبکه خروجی GRN محاسبه می‌شود. چهار الگوریتم از بین روش‌های شناخته شده و جدید برای استنتاج شبکه تنظیم بیان ژن انتخاب شده است. در ادامه توضیح مختصری از عملکرد آن‌ها آمده است.

Genie3: در این روش از Random Forest و Extra-Trees برای ایجاد برنامه تنظیمی هر ژن هدف استفاده می‌شود. در هر درخت رگرسیون، تنظیم‌کننده‌ها به‌عنوان ویژگی مورداستفاده قرار می‌گیرند و بیان هر ژن هدف به‌عنوان تابعی غیرخطی از بیان تنظیم‌کننده‌ها مدل می‌شود [۲۸].

TIGRESS: در این روش از رگرسیون LASSO در یک چارچوب انتخاب پایداری برای استنباط برنامه تنظیمی هر ژن هدف استفاده می‌شود [۲۹].

RGBM: یک چارچوب جدید استنتاج GRN است که مدل اصلی آن برای استنباط روابط تنظیمی برای هر ژن هدف می‌تواند boosting of regression stumps یا مجموعه درختان تصمیم‌گیری (Random-Forests) RF باشد [۳۰].

GENIMS: این روش یک استراتژی سه سطحی را اتخاذ می‌کند. سطح اول حل مسئله رگرسیون انفرادی با الگوریتم جنگل تصادفی هدایت شده است، سطح دوم اعمال نرمال‌سازی q -norm برای کاهش بایاس در بین نتایج مختلف رگرسیون است و سطح سوم اصلاح نتایج قبلی با توجه به خاصیت اسپارسیتی شبکه‌های تنظیمی ژن در مقیاس بزرگ هست [۳۱].

برای بررسی کیفیت ویژگی‌های حاصل از روش پیشنهادی، هر کدام از روش‌های استنتاج اشاره شده در بالا، با داده‌های ورودی مختلف اجرا شده‌اند. یک بار با ویژگی‌های داده بیان ژن به تنهایی و یک بار هم با ویژگی‌های به دست آمده از روش پیشنهادی. در جدول ۱ و ۲ نتایج میانگین ۵ بار اجرای هرکدام از روش‌های استنتاج هنگامی‌که ورودی تنها نوع داده بیان ژن می‌باشد و همین‌طور هنگامی‌که ورودی بردار ویژگی حاصل از ادغام داده‌ها می‌باشد برای مجموعه داده‌های Saccharomyces cerevisiae و Escherichia coli آورده شده است. مشهود است که استفاده از ویژگی‌های حاصل از روش پیشنهادی منجر به بهبود استنتاج همه روش‌ها در معیار AUROC ناحیه زیر منحنی (Receiver Operating Characteristic) ROC و AUPR (ناحیه زیر منحنی دقت–حساسیت) شده است.

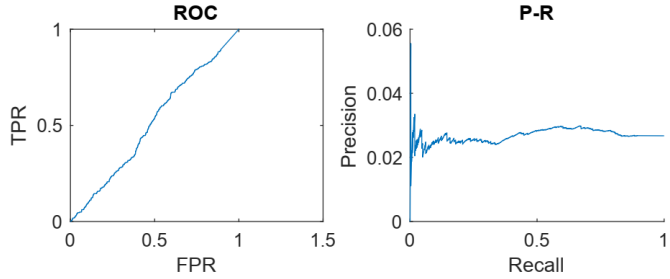
جدول ۱: نتایج میانگین ۵ بار اجرای الگوریتم‌های استنتاج برای مجموعه داده Saccharomyces cerevisiae

داده	داده بیان ژن		داده ادغامی پیشنهادی	
روش	AUROC	AUPR	AUROC	AUPR
Genie3	۰/۵۰۳۶	۰/۰۲۶۵	۰/۵۵۰۰	۰/۰۳۴۰
TIGRESS	۰/۵۰۲۱	۰/۰۲۶۹	۰/۵۳۱۳	۰/۰۳۲۷
RGBM	۰/۵۰۴۸	۰/۰۲۸۲	۰/۵۱۷۵	۰/۰۳۱۱
GENIMS	۰/۵۱۳۱	۰/۰۲۷۳	۰/۵۴۲۳	۰/۰۳۳۱

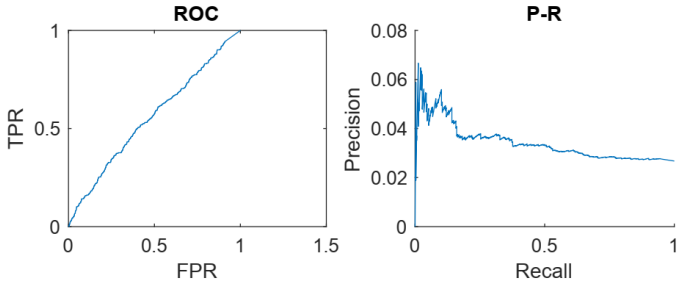
جدول ۲: نتایج میانگین ۵ بار اجرای الگوریتم‌های استنتاج برای مجموعه داده E.coli

داده	داده بیان ژن		داده ادغامی پیشنهادی	
روش	AUROC	AUPR	AUROC	AUPR
Genie3	۰/۶۰۳۸	۰/۰۶۸۲	۰/۶۷۴۷	۰/۱۲۶۷
TIGRESS	۰/۵۶۳۳	۰/۰۵۶۳	۰/۵۷۵۶	۰/۰۶۴۰
RGBM	۰/۵۷۴۲	۰/۰۵۶۵	۰/۶۱۳۱	۰/۱۱۴۵
GENIMS	۰/۵۹۰۳	۰/۰۵۰۱	۰/۶۴۶۵	۰/۱۰۵۱

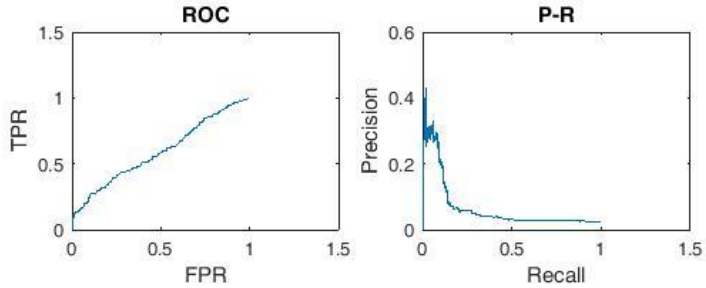
نمودارهای AUROC و AUPR برای نتایج به دست آمده (جدول ۱ و ۲) برای روش استنتاج Genie3 هنگامی‌که ورودی تنها نوع داده بیان ژن می‌باشد و همین‌طور هنگامی‌که ورودی بردار ویژگی حاصل از ادغام داده‌ها می‌باشد برای مجموعه داده Saccharomyces cerevisiae و coli Escherichia در شکل‌های ۳ تا ۶ آمده است.



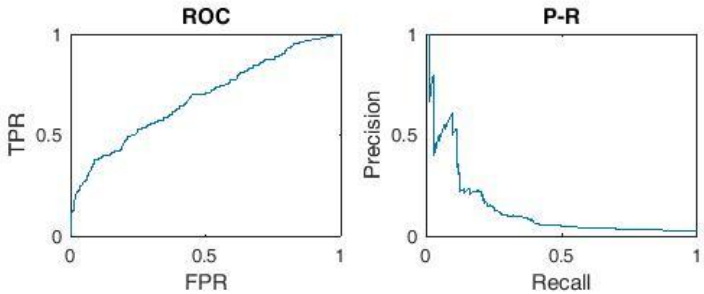
شکل ۳: نمودارهای AUROC و AUPR برای روش Genie3 و نوع داده بیان ژن برای مجموعه داده Saccharomyces cerevisiae



شکل ٤: نمودارهای AUROC و AUPR برای روش Genie3 و نوع داده ادغامی برای مجموعه داده Saccharomyces cerevisiae



شکل ٥: نمودارهای AUROC و AUPR برای روش Genie3 و نوع داده بیان ژن برای مجموعه داده Escherichia coli



شکل ٦: نمودارهای AUROC و AUPR برای روش Genie3 و نوع داده ادغامی برای مجموعه داده Escherichia coli

برای مقایسه روش پیشنهادی با سایر روش‌ها، روش‌های [۷] SNF و SVD (Singular Value Decomposition) انتخاب شده است. برای روش SNF که مبتنی بر شبکه می‌باشد، الگوریتم پایه SNF تغییر داده شد تا قابلیت پذیرش داده‌های پیشنهادی را داشته باشد. در کنار این روش الگوریتم Singular Value Decomposion نیز به عنوان یک

روش استخراج ویژگی برای مقایسه استفاده شده است. نتایج این مقایسه بین روش پیشنهادی با روش SVD و روش SNF در جدول ۳ و ۴ آمده است.

جدول ۳: مقایسه بین روش پیشنهادی با روش SVD و روش SNF براساس معیار AUROC و AUPR در مجموعه داده S. cerevisiae

روش	روش پیشنهادی		Singular Value Decomposition		Similarity Network Fusion	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Genie3	**۰/۵۵۰۰**	۰/۰۳۴۰	۰/۵۱۶۴	۰/۰۴۱۳	۰/۵۴۲۰	۰/۰۳۳۰
TIGRESS	**۰/۵۳۱۳**	۰/۰۳۲۷	۰/۴۹۹۲	۰/۰۳۷۸		
RGBM	۰/۵۱۵۷	۰/۰۳۱۱	۰/۴۸۲۷	۰/۰۲۹۰		
GENIMS	**۰/۵۴۲۳**	**۰/۰۳۳۱**	۰/۵۱۰۰	۰/۰۲۸۰		

جدول ٤: مقایسه بین روش پیشنهادی با روش SVD و روش SNF براساس معیار AUROC و AUPR در مجموعه داده E.coli

روش	روش پیشنهادی		Singular Value Decomposion		Similarity Network Fusion	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Genie3	**۰/٦٧٤٧**	۰/۱۲۶۷	۰/۶۰۱۱	۰/۱۵۱۱	۰/۵۳۷۸	۰/۰۵۱۷
TIGRESS	**۰/٥٧٥٦**	**۰/۰٦٤۰**	۰/۴۸۸۸	۰/۰۳۴۳		
RGBM	**۰/٦۱۳۱**	**۰/۱۱٤٥**	۰/۵۲۸۵	۰/۰۷۰۹		
GENIMS	**۰/٦٤٦٥**	**۰/۱۰٥۱**	۰/۵۵۷۵	۰/۰۶۵۲		

بحث و نتیجه‌گیری

آنچه در این پژوهش دنبال شده است، بر اساس چالش یکپارچه‌سازی داده‌ها برای کمک به بهبود استنتاج شبکه‌های تنظیم بیان ژن بود. فرآیند تنظیم بیان ژن، یک فرآیند چند لایه‌ای است، از رونویسی تا سنتز پروتئین؛ اما اکثر مطالعات در حوزه استنتاج شبکه‌های ژنی تنها روی داده بیان ژن متمرکز شده‌اند و مکانیزم‌های رونویسی به عنوان داده اصلی برای محاسبات ژنتیکی سیستماتیک محسوب می‌شود. این در حالی است که بازسازی این شبکه‌ها تنها از روی یک نوع داده نمی‌تواند دقیق باشد و درک مفهومی سیستم بیولوژیک می‌تواند تنها با تحلیل ارتباط همه لایه‌های اُمیکس به دست آید. در این پژوهش، یک چارچوب تحلیلی پیشنهاد شد که ترکیبی از روش مبتنی بر انتشار با کاهش ابعاد برای استخراج بهتر اطلاعات توپولوژیکی شبکه‌های مختلف ورودی از لایه‌های امیکس متفاوت به منظور تسهیل شناسایی ارتباط ژن‌ها است. ایده اصلی به دست آوردن ویژگی‌های حاوی اطلاعات مفید؛ اما با ابعاد کم است که خصوصیات توپولوژیکی ذاتی هر گره در شبکه را در بر می‌گیرد. ابتدا الگوریتم گام تصادفی با راه‌اندازی مجدد در هر شبکه داده‌ای ورودی اجرا می‌شود تا توزیع هر گره که ارتباط آن گره با سایر گره‌های شبکه است به دست آید. سپس هر یک از این توزیع‌ها با یک مدل لجستیک چند جمله‌ای برای هر گره تقریب زده می‌شود. بردارهای ویژگی برای هر گره با به حداقل رساندن واگرایی (KL) (آنتروپی نسبی) [۲۳] بین توزیع‌های هر گره و توزیع لجستیک پارامتری چندجمله‌ای به دست می‌آیند. علاوه بر این، این چهارچوب قابلیت گسترش برای ادغام شبکه‌های ناهمگن متعدد را به صورت انتشار در شبکه‌های جداگانه و بهینه‌سازی بردارهای ویژگی به طور مشترک را دارا است. نهایتاً از این بردارهای ویژگی به عنوان ورودی برای الگوریتم‌های استنتاج شبکه‌های تنظیم بیان ژن استفاده می‌شود. مطالعه موردی برای شبکه‌های Saccharomyces cerevisiae و Escherichia coli نشان می‌دهد که ادغام اطلاعات مبتنی بر الگوریتم انتشار در مقایسه با داده بیان ژن به تنهایی، منجر به بهبود ۰/۰۱ الی ۰/۰۷ واحد در معیار AUROC در روش‌های مختلف استنتاج شبکه‌های ژنی می‌شود.

در پژوهش Wang و همکاران [۷] SNF که داده بیان mRNA, miRNA و داده متیلاسیون DNA را ادغام می‌کند، یک شبکه یکپارچه کلی از ادغام همه شبکه‌های همگن تولید می‌کند، اما یک محدودیت اساسی در چنین رویکردهایی، از دست دادن اطلاعات با نگاشت مجموعه

داده‌های مختلف روی یک شبکه واحد می‌باشد؛ که در این پژوهش این چالش با اجرای الگوریتم انتشار به صورت جداگانه بر شبکه‌های ورودی حل شده است. از عوامل دیگر موفقیت روش پیشنهادی، ابعاد کم و فشردگی بردار ویژگی تولیدی آن است که به جداسازی الگوهای توپولوژیکی و عملکردی از نویز در داده‌ها کمک می‌کند که موجب می‌شود که عملکرد بهتری نسبت به الگوریتم پایه‌ای همچون SVD از خود نشان می‌دهد.

بهبود استنتاج به‌دست‌آمده از روش پیشنهادی را می‌توان به این علت دانست که تجزیه و تحلیل جداگانه ساختار هر شبکه، از الگوهای توپولوژیکی مخفی‌ای رونمایی می‌کند که کشف آن‌ها در یک شبکه ترکیبی که انواع یال‌های مختلف از هم متمایز نیستند، دشوار است. به‌عنوان‌مثال ژن‌هایی که بر اساس بردار ویژگی پیشنهادی، از نظر توپولوژیک مشابه هستند، ممکن است در هیچ‌یک از شبکه‌ها همسایه مستقیم نباشند، بلکه ممکن است با ژنی مشابه باشد که به‌طور غیرمستقیم با مسیرهای متعدد از طریق گره‌های میانی مختلف متصل می‌شوند و ارتباط غیرمستقیم دارند. چنین ارتباطات غیرمستقیم؛ اما محکم و استواری اگر در یک شبکه ترکیبی واحد تحلیل شوند، اغلب توسط همسایگان مستقیم تحت‌الشعاع و نادیده قرار می‌گیرند حتی اگر اتصالات مستقیم فقط در یک زمینه محدود و در شبکه‌های معدودی وجود داشته باشند.

از محدودیت‌های روش پیشنهادی اول این که، با توجه به دانش ناقص ما در مورد شبکه‌های بیولوژیکی، این که چه اندازه ویژگی‌های محاسبه شده به کشف اتصالات کمک می‌کند، طبیعتاً به کیفیت و کامل بودن شبکه‌های ارائه شده ورودی محدود می‌شود. دوم این که محدودیت‌های مربوط به شبکه‌های معیار و استاندارد مانع از کشف و بررسی اتصالات جدید می‌شود. البته این مسئله هم باید در نظر گرفته شود که در صورت کشف اتصالات جدید حتماً باید به طور تجربی هم آزمایش شوند. محدودیت دیگر روش نوع داده ورودی است که به صورت شبکه می‌باشد.

برای بهبود روش پیشنهادی، ادغام انواع داده بیولوژی غیر شبکه‌ای همچون داده‌های توالی ژنوم پیشنهاد می‌شود. برای پژوهش‌های آتی می‌توان روش پیشنهادی را بر روی مجموعه داده‌های بیولوژیکی یا سرطانی بیشتری آزمایش کرد. بررسی کارایی بردارهای ویژگی تولید شده برای حل مسئله‌های بیولوژیکی دیگر هم می‌تواند مسیرهای پژوهشی متنوعی را به وجود آورد.

تعارض منافع

نویسندگان با یکدیگر تعارض منافع نداشتند.

References

1. Pevsner J. Bioinformatics and Functional Genomics. 3rd ed. UK: Wiley-Blackwell; 2015.

2. Lodish H, Berk A, Kaiser CA, Krieger M, Scott MP, Bretscher A, et al. 4th ed. Molecular Cell Biology. New York: Macmillan; 2008.

3. Sanguinetti G, Huynh-Thu VA. Gene Regulatory Networks: Methods and Protocols. 1st ed. New York: Humana; 2019.

4. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-comics data-a review. Front Genet. 2019; 10: 535. doi: 10.3389/fgene.2019.00535

5. Banf M, Rhee SY. Computational inference of gene regulatory networks: approaches, limitations and opportunities. Biochim Biophys Acta Gene Regul Mech 2017;1860(1):41-52. doi: 10.1016/j.bbagrm.2016.09.003

6. Angelin-Bonnet O, Biggs PJ, Vignes M. Gene regulatory networks: a primer in biological processes and statistical modelling. Methods Mol Biol 2019;1883:347-83. doi: 10.1007/978-1-4939-8882-2_15

7. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333-7. doi: 10.1038/nmeth.2810

8. Wong AK, Park CY, Greene CS, Bongo LA, Guan Y, Troyanskaya OG. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Res 2015; 43(Web Server issue): W128–33. doi: 10.1093/nar/gkv486

9. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol 2008;9 Suppl 1(Suppl 1):S4. doi: 10.1186/gb-2008-9-s1-s4

10. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 2013;41(Database issue):D808-15. doi: 10.1093/nar/gks1094

11. Isci S, Dogan H, Ozturk C, Otu HH. Bayesian network prior: network analysis of biological data using external knowledge. Bioinformatics 2014;30(6):860-7. doi: 10.1093/bioinformatics/btt643

12. Santra T. A bayesian framework that integrates heterogeneous data for inferring gene regulatory

networks. Front Bioeng Biotechnol 2014; 2: 13. doi: 10.3389/fbioe.2014.00013
13. Žitnik M, Zupan B. Gene network inference by fusing data from diverse distributions. Bioinformatics 2015;31(12):i230-9. doi: 10.1093/bioinformatics/btv258
14. Banf M, Rhee SY. Enhancing gene regulatory network inference through data integration with markov random fields. Sci Rep 2017;7:41174. doi: 10.1038/srep41174
15. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics 2004;20(16):2626-35. doi: 10.1093/bioinformatics/bth294
16. Yamanishi Y, Vert JP, Kanehisa M. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics 2004;20(suppl_1):i363-70. doi: 10.1093/bioinformatics/bth910
17. Žitnik M, Zupan B. Data fusion by matrix factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 2014;37(1):41-53. doi: 10.1109/TPAMI.2014.2343973
18. Hwang T, Atluri G, Xie M, Dey S, Hong C, Kumar V, Kuang R. Co-clustering phenome–genome for phenotype classification and disease gene discovery. Nucleic Acids Res 2012;40(19):e146. doi: 10.1093/nar/gks615
19. Žitnik M, Janjić V, Larminie C, Zupan B, Pržulj N. Discovering disease-disease associations by fusing systems-level molecular data. Sci Rep 2013;3:3202. doi: 10.1038/srep03202
20. Cho H, Berger B, Peng J. Diffusion Component Analysis: Unraveling Functional Topology in Biological Networks. Res Comput Mol Biol. 2015;9029:62-4. doi: 10.1007/978-3-319-16706-0_9
21. Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, Cau P, Remy E, Baudot A. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics 2019;35(3):497-505. doi: 10.1093/bioinformatics/bty637
22. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 2017;8(1):573. doi: 10.1038/s41467-017-00680-8
23. Van Erven T, Harremos P. Rényi divergence and Kullback-Leibler divergence. IEEE Transactions on Information Theory 2014;60(7):3797-820. doi: 10.1109/TIT.2014.2320500
24. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS) 1997;23(4):550-60. https://doi.org/10.1145/279232.279236
25. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47(D1):D607-13. doi: 10.1093/nar/gky1131
26. Orlando DA, Lin CY, Bernard A, Wang JY, Socolar JE, Iversen ES, Hartemink AJ, Haase SB. Global control of cell-cycle transcription by coupled CDK and network oscillators. Nature 2008;453(7197):944-7. doi: 10.1038/nature06955
27. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. Nat Methods 2012; 9(8): 796–804. doi: 10.1038/nmeth.2016
28. Huynh-Thu VA. Machine learning-based feature ranking: Statistical interpretation and gene network inference [dissertation]. Belgium: Université de Liège, Liège; 2012.
29. Haury AC, Mordelet F, Vera-Licona P, Vert JP. TIGRESS: trustful inference of gene regulation using stability selection. BMC Syst Biol 2012;6:145. doi: 10.1186/1752-0509-6-145
30. Mall R, Cerulo L, Garofano L, Frattini V, Kunji K, Bensmail H, et al. RGBM: regularized gradient boosting machines for identification of the transcriptional regulators of discrete glioma subtypes. Nucleic Acids Res 2018;46(7):e39. doi: 10.1093/nar/gky015
31. Wu J, Zhao X, Lin Z, Shao Z. Large scale gene regulatory network inference with a multi-level strategy. Mol Biosyst 2016;12(2):588-97. doi: 10.1039/c5mb00560d

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2020; 7(2): 201-213

Improving the Inference of Gene Expression Regulatory Networks with Data Aggregation Approach

Naseri Atefeh[1], Hasheminejad Seyed Mohammad Hossein[2]*, Sharghi Mehran[3]

• **Received:** 23 Sep 2019 • **Accepted:** 31 May 2020

Introduction: The major issue for the future of bioinformatics is the design of tools to determine the functions and all products of single-cell genes. This requires the integration of different biological disciplines as well as sophisticated mathematical and statistical tools. This study revealed that data mining techniques can be used to develop models for diagnosing high-risk or low-risk lifestyles for colorectal cancer.

Method: In this retrospective study, a dataset consisting of information relevant to 84 patients and 225 healthy individuals with 25 attributes was collected. This information was on patients diagnosed from 2006 to the first quarter of 2014. The most widely used techniques in the medical informatics literature including support vector machine, Naive Bayes, decision tree, and k-nearest neighbor were used to develop the models.

Results: The developed models are able to distinguish people's lifestyles efficiently. A well-developed non-technical measure can properly determine the true value of individual predictions, whether true or false, at actual costs, and indicate a true measure of the cost savings in the health system by each model. Among the developed models, only two models were able to meet the criteria set for use in the real world.

Conclusion: The developed models should not only be technically evaluated, but should also be examined in terms of metrics accepted for the medical field as well as feasibility for real problem solving.

Keywords: Gene Expression Regulatory Network, Gene Expression Regulatory Network Inference, Propagation Algorithm, Data integration

• **Citation:** Naseri A, Hasheminejad SM, Sharghi M. Improving the Inference of Gene Expression Regulatory Networks with Data Aggregation Approach. Journal of Health and Biomedical Informatics 2020; 7(2): 201-13. [In Persian]

1. M.Sc. in Computer Engineering, Computer Engineering Dept., Faculty of Engineering, Alzahra University, Tehran, Iran
2. Ph.D. in Computer Engineering, Assistant Professor, Computer Engineering Dept., Faculty of Engineering, Alzahra University, Tehran, Iran
3. Ph.D. in Computer Engineering, Assistant Professor, Computer Engineering Dept., Faculty of Engineering, Alzahra University, Tehran, Iran

***Correspondence**: Seyed Mohammad Hossein Hasheminejad
Address: Vanak Square, Alzahra University, Tehran, Iran
• **Tel:** 021-88044051 - 2170 • **Email:** SMH.Hasheminejad@Alzahra.ac.ir