مجله دانشکده پزشکی اصفهان
سال سی و سوم/شماره ۳۵۹/هفته سوم دی ماه ۱۳۹٤

تاریخ دریافت: ۱۳۹۴/۷/۱۰
تاریخ پذیرش: ۱۳۹۴/۹/۵

بررسی تأثیر کاهش بعد بر فضای ویژگی‌های توپولوژیک شبکه‌ی ژن ساخته شده از میکروآرایه به منظور پیش‌گویی عود سرطان سینه

دکتر علیرضا مهری دهنوی[۱]، حامد زند[۲]، دکتر محمدرضا صحتی[۳]

مقاله پژوهشی

چکیده

مقدمه: یکی از روش‌های متداول در طبقه‌بندی نمونه‌های سرطانی، استفاده از ویژگی‌های به دست آمده از داده‌ی بیان ژن در میکروآرایه‌های DNA می‌باشد. در این خصوص، با استفاده از ویژگی‌های توپولوژیک شبکه‌ی ژن بازسازی شده از داده‌های بیان ژن، می‌توان با بهره گرفتن از اطلاعات تعامل بین ژن‌ها، به یافته‌های مطمئن‌تری دست یافت. هدف از انجام این مطالعه، پیش‌گویی عود سرطان سینه بر اساس انتخاب ویژگی مبتنی بر ویژگی‌های توپولوژیک، متناظر با شبکه‌ی ارتباطی ژن‌ها بود.

روش‌ها: هفت مجموعه داده‌ی بیان ژن میکروآرایه شامل ۱۲۷۱ نمونه مربوط به سرطان سینه در مطالعه‌ی حاضر مورد استفاده قرار گرفت. ابتدا شبکه‌ی ارتباطی ژن‌ها از داده‌های آموزش با اعمال روش انتخاب ویژگی Fisher (Fisher discriminant analysis) بر داده‌های ویژگی توپولوژیک این شبکه، ساخته شد. به دلیل این‌که نمی‌توان برای یک نمونه شبکه‌ی ژن ساخت؛ این نمونه به کل داده‌های آموزشی اضافه و دوباره شبکه‌ی ژن ساخته شد. سپس، همبستگی بین بردارهای ویژگی توپولوژیک ژن‌های شاخص در دو شبکه، قبل و بعد از اضافه شدن نمونه‌ی آزمایش محاسبه گردید. در نهایت، نمونه‌ای با همبستگی بیشتر در یک کلاس نسبت به کلاس دیگر، جزء همان کلاس در نظر گرفته شد.

یافته‌ها: صحت پیش‌گویی به دست آمده بر اساس ویژگی‌های توپولوژیک مربوط به شبکه‌ی بازسازی شده از داده‌های بیان ژن نسبت به انتخاب ویژگی مستقیم از این داده‌ها بالاتر بود. بیشترین صحت طبقه‌بندی بر اساس ویژگی توپولوژیک توزیع درجه‌ی تمرکز (متوسط ۹۸/۵ درصد در داده‌ی آزمایش) به دست آمد.

نتیجه‌گیری: استفاده از ساختار توپولوژی شبکه‌ی ژنی، ویژگی‌های پایدارتری را نسبت به کاربرد مستقل مقدار بیان ژن به منظور پیش‌گویی و طبقه‌بندی سرطان فراهم می‌آورد.

واژگان کلیدی: سرطان سینه، بیان ژن، شبکه‌ی ژن، توپولوژی

ارجاع: مهری دهنوی علیرضا، زند حامد، صحتی محمدرضا. **بررسی تأثیر کاهش بعد بر فضای ویژگی‌های توپولوژیک شبکه‌ی ژن ساخته شده از میکروآرایه به منظور پیش‌گویی عود سرطان سینه.** مجله دانشکده پزشکی اصفهان ۱۳۹۴؛ ۳۳ (۳۵۹): ۱۹۸۵-۱۹۷۴

مقدمه

سرطان سینه یکی از سرطان‌های شایع در میان جمعیت زنان جهان می‌باشد. بر اساس آخرین آمار مرکز تحقیقات سرطان ایران، سالانه حدود ۸۵۰۰ مورد جدید سرطان پستان در کشور ثبت می‌شود و ۱۴۰۰ نفر به دلیل ابتلا به سرطان پستان فوت می‌کنند. همچنین، در حال حاضر حدود ۴۰۰۰۰ بیمار مبتلا به این بیماری در کشور زندگی می‌کنند. در بین انواع مختلف سرطان، سرطان سینه ۲۳ درصد همه‌ی سرطان‌ها در زنان را شامل می‌شود (۱). بنابراین، تشخیص به‌هنگام و دقیق‌تر مراحل و میزان پیشرفت این سرطان امری مهم و حیاتی محسوب می‌گردد. در بیشتر موارد، این تشخیص در محیط آزمایشگاه و طبق نظر پاتولوژیست صورت می‌گیرد. در مرحله‌ی بعد، در صورت پی بردن به سرطانی بودن بافت، مرحله‌ی پیشرفت آن تعیین و در نهایت بازگشت‌پذیر بودن (۳-۲) یا عود سرطان در آینده مشخص خواهد شد، اما به منظور افزایش صحت تصمیم‌گیری، بحث شکل‌گیری و توسعه‌ی سرطان باید در سطح ژنوم انسان مورد بررسی قرار گیرد (۴). به همین منظور، می‌توان از فن‌آوری میکروآرایه برای

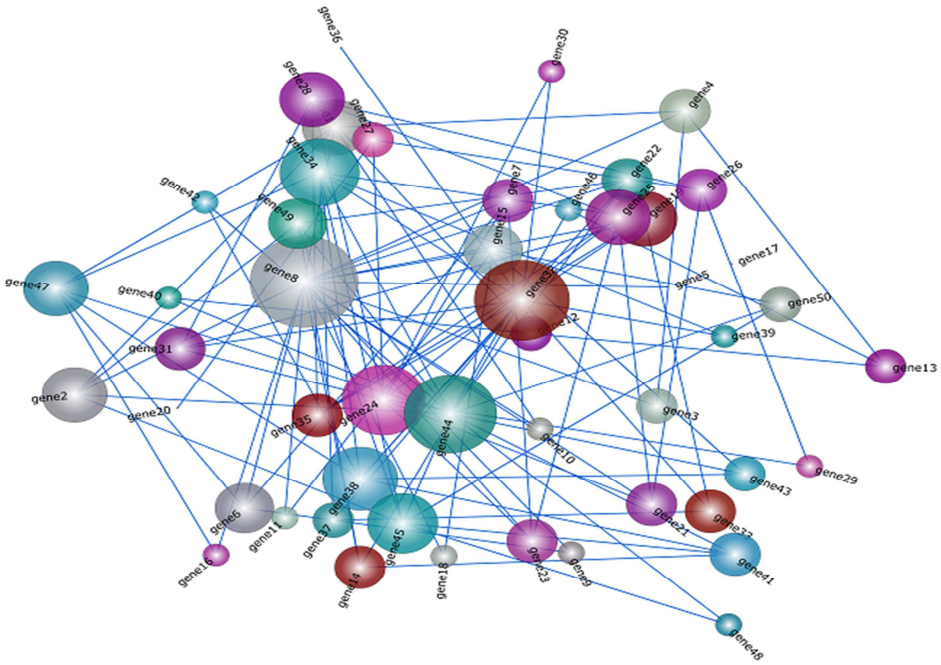
۱- دانشیار، گروه بیوالکتریک، دانشکده‌ی فن‌آوری‌های نوین علوم پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۲- دانشجوی کارشناسی ارشد، گروه بیوالکتریک، دانشکده‌ی فن‌آوری‌های نوین علوم پزشکی و کمیته‌ی تحقیقات دانشجویی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۳- استادیار، گروه بیوالکتریک، دانشکده‌ی فن‌آوری‌های نوین علوم پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
نویسنده‌ی مسؤول: دکتر محمدرضا صحتی
Email: mr.sehhati@gmail.com

بررسی بیان ژن‌ها در بافت تومور استفاده نمود (۵).

با استفاده از فن‌آوری میکروآرایه، مجموعه ژن‌هایی از بافت مشخصی مانند بافت سینه به طور هم‌زمان از نظر سطح بیان ژن‌ها مورد بررسی قرار می‌گیرند (۶)، اما مقدار بیان اندازه‌گیری شده برای هر ژن توسط میکروآرایه، اطلاعی از نحوه‌ی تعامل آن ژن با سایر ژن‌ها ارایه نخواهد داد. برای رفع این کمبود، باید از مدل‌های تعاملی یا شبکه‌های بیولوژیک استفاده کرد که در آن‌ها تعامل بین ژن‌ها و پروتئین‌ها نیز در نظر گرفته می‌شود. مدل ریاضی این شبکه‌ی تعاملی را می‌توان به صورت یک گراف در نظر گرفت که گره‌های آن همان ژن‌ها هستند و یال‌های این گراف، وجود میان‌کنش بین دو ژن (گره) را نمایش می‌دهند. شبکه‌های بیولوژیک استاندارد تحت شرایط آزمایشی خاصی ساخته می‌شوند و با وجود این‌که داده‌های ارزشمند و مفیدی در اختیار ما قرار می‌دهند، اما در مجموع داده‌های کاملی نیستند و عاری از خطا نمی‌باشند (۷). در این باره یکی از استراتژی‌هایی که محققان در حال حاضر دنبال می‌کنند، ساخت شبکه‌های ژنی است که تعاملات بین ژن‌ها را بر اساس تغییرات سطح بیان آن‌ها مورد بررسی قرار می‌دهد. نمونه‌ی ساده از شبکه‌ی تعامل ژن با ژن در شکل ۱ نشان داده شده است (۸).

با ساخت شبکه‌ی ژن از روی مجموعه داده‌های بیان ژن در سطح میکروآرایه، می‌توان علاوه بر اطلاعات مستقل ژن‌ها، اطلاعاتی در مورد تعاملات و تأثیرگذاری ژن‌های مختلف نسبت به یکدیگر کسب نمود تا از این طریق درک بهتری از وضعیت نمونه‌های مورد بررسی حاصل شود. جهت استنتاج شبکه‌ی ژن از روی داده‌های بیان ژن، روش‌های مختلفی وجود دارد که از جمله مهم‌ترین آن‌ها می‌توان به مدل‌های ارتباطی (Relevance network) (۹)، مدل‌های منطقی (Logic models) (۱۰)، مدل‌های Bayesian (Bayesian models) (۱۱)، معادلات تفاضل (Differential equations) (۱۲)، مدل فضای حالت (State space model) (۱۳)، مدل‌های مبتنی بر شبکه‌ی عصبی (Neural network) (۱۴) و مدل‌های پیمانه (Modular models) (۱۵) اشاره نمود.

به منظور ساخت شبکه در مطالعه‌ی حاضر و به دلیل ساده بودن و حجم پایین محاسبات و همچنین، تناسب بیشتر مجموعه‌ی داده‌ها برای مدل‌سازی شبکه‌ی ارتباطی، مدل نوع ارتباطی نسبت به مدل‌های دیگر جهت ساخت شبکه‌ی ژن ارجحیت داده شد. در عمل از یک نمونه نمی‌توان یک شبکه‌ی ژن ساخت؛ چرا که به مجموعه‌ای از نمونه‌ها نیاز است تا استنتاج شبکه‌ی ژن محقق شود. از طرف دیگر، نمی‌توان در مورد نمونه‌ای که از آن در ساخت شبکه استفاده نشده است، بر اساس نمونه‌های استفاده شده تصمیم‌گیری نمود. در مطالعه‌ی Liu و همکاران طبقه‌بندی هر نمونه‌ی سرطان بر اساس ساخت شبکه‌ی ژن از کل داده‌ها و اضافه شدن نمونه‌ی آزمایش به هر گروه داده‌ها بر اساس محاسبه‌ی ضریب همبستگی Pearson بین شبکه‌ها انجام شد (۱۶).



شکل ۱. شبکه‌ی تعامل ژن با ژن

در مطالعه‌ی دیگری از Liu و همکاران، آنالیز مسیرهای بیولوژیک (Pathway analysis) با ساخت شبکه‌ی ژن مبتنی بر ضریب رتبه‌بندی ژن‌ها به جای شبکه‌های مبتنی بر همبستگی بین ژن‌ها صورت گرفت (۱۷).

Raza و Jaiswal با ساخت شبکه‌های تنظیم ژن از داده‌های بیان ژن، روابط تنظیم کننده‌ی ژن‌ها در شبکه‌های تنظیم ژن را بررسی کردند که برخی از ژن‌های شبکه‌ی تنظیم از نظر بیان بیشتر در تمامی نمونه‌ها به عنوان هدف‌های دارویی جهت تشخیص سرطان در نظر گرفته شدند (۱۸). تحقیق Yang و همکاران چارچوب استنتاج مبتنی بر شبکه‌ی تفاضلی را [که یک روش آنالیز آماری مبتنی بر شبکه (Network-based statistical analysis method) است] به دو مجموعه داده‌ی مستقل سرطان سینه اعمال کردند که با تفاضل دو شبکه‌ی ژن از دو مجموعه داده‌ی سرطان، مدل شبکه‌ی تفاضلی به دست آمد. در نهایت از شبکه‌ی تفاضلی در رتبه‌بندی ژن‌ها برای شناسایی بیومارکرها استفاده شد (۱۹).

پژوهش Chuang و همکاران، روشی مبتنی بر شبکه‌های پروتئینی ارایه نمود که از نشانگرهایی به عنوان زیرشبکه‌های استخراج شده از داده‌های تعامل پروتئینی در طبقه‌بندی سرطان استفاده کرده بود (۲۰). Ahn و همکاران با استفاده از داده‌های نرمالیزه شده‌ی سرطان پروستات، شبکه‌های ژنی برای کلاس نرمال و سرطانی را بازسازی کردند. آن‌ها در مطالعه‌ی خود از شبکه‌های ژن استاندارد و آنالیز مسیر بیولوژیک استفاده نمودند (۲۱). در مطالعه‌ی Bockmayr و همکاران، الگوریتم‌های DCloc و DCglob توسعه یافت و با مقایسه‌ی تغییرات توپولوژیک از شبکه‌های ژن ساخته شده از دو بیماری مختلف، الگوهای همبستگی تفاضلی بین شبکه‌های ساخته شده از دو بیماری شناسایی شد (۲۲).

هدف اصلی از انجام مطالعه‌ی حاضر، کاهش بعد یا حذف ژن‌های غیر مفید در طبقه‌بندی سرطان سینه بر اساس ساخت شبکه‌ی ژن و استفاده از انتخاب ویژگی از روی داده‌های ویژگی توپولوژیک شبکه‌ی ژن بود. در بیشتر مطالعات قبلی، انتخاب ژن به طور مستقیم از روی داده‌ی بیان ژن انجام شده بود؛ به طوری که ساخت شبکه‌ی ژن از روی داده‌ی بیان ژن صورت می‌گرفت، اما در تحقیق حاضر برای ساخت شبکه‌ی ژن از داده‌های ویژگی توپولوژیک استفاده گردید. لازمه‌ی ساخت شبکه‌ی ژن از داده‌ی توپولوژیک، استخراج داده‌های ویژگی توپولوژیک بعد از ساخت شبکه‌ی ژن از داده‌های بیان ژن در میکروآرایه می‌باشد. انتخاب بهترین ویژگی توپولوژی شبکه در جهت دستیابی به بالاترین صحت پیش‌گویی عود سرطان سینه و نیز شناسایی ژن‌های شاخص مرتبط با عود سرطان، در مطالعه‌ی حاضر مورد بررسی قرار گرفت.

روش‌ها

تهیه‌ی داده‌ها: مجموعه داده‌ی مورد استفاده، مربوط به هفت مطالعه‌ی مستقل در زمینه‌ی سرطان سینه و شامل ۱۲۷۱ نمونه بود که از سایت مرکز ملی اطلاعات بیوتکنولوژی (National Center of Biotechnology Information) دانلود شد. قبل از ساخت شبکه‌ی ژن و به منظور نرمال کردن داده‌ها، ابتدا تابع لگاریتم بر داده‌های میکروآرایه اعمال شد و سپس، مقادیر بیان ژن در راستای ستون‌ها و سطرهای ماتریس به طور متوالی نرمال گردید. داده‌ها شامل دو گروه سرطانی کم‌خطر (نمونه‌هایی که تا پنج سال از زمان تشخیص در آن‌ها متاستاز رخ نداده است) و پرخطر (نمونه‌هایی که در فاصله‌ی کمتر از پنج سال از زمان تشخیص در آن‌ها متاستاز رخ داده است) بود و در مجموع، ۸۹۲ نمونه‌ی کم‌خطر و ۳۷۹ نمونه‌ی پرخطر بررسی گردید. از گروه پرخطر ۸۰ نمونه به طور تصادفی به عنوان نمونه‌ی آزمایش و ۲۹۹ نمونه‌ی باقی‌مانده به عنوان داده‌ی آموزشی و در گروه کم‌خطر نیز ۱۸۰ نمونه برای آزمایش و ۷۱۲ نمونه برای آموزش در نظر گرفته شد. ساخت شبکه‌ی ژن از روی داده‌های آموزشی برای هر دو گروه داده‌ها به طور مستقل انجام گرفت.

انتخاب ژن مبتنی بر داده‌ی ویژگی توپولوژیک: در تمامی مراحل بررسی از جمله در مرحله‌ی ساخت شبکه‌ی ژن، از روش انتخاب ویژگی مبتنی بر فیلتر Fisher (Fisher discriminant analysis) (۲۳) استفاده گردید که تابع امتیازدهی S در آن به صورت رابطه‌ی ۱ است.

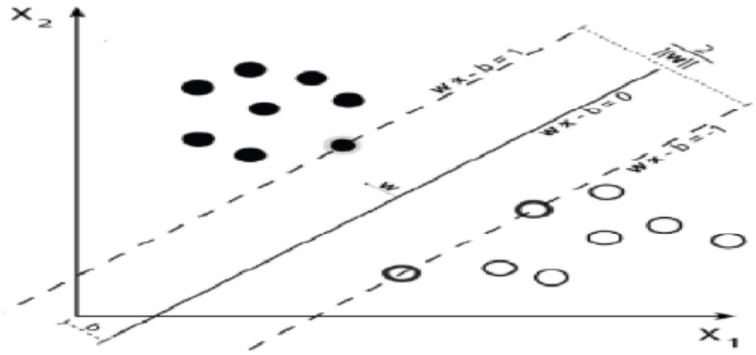
$$S_{12}(f) = \frac{|m1-m2|}{(\sigma1+\sigma2)} \qquad \text{رابطه‌ی ۱}$$

عنصر m بیانگر میانگین بیان ژن‌ها در نمونه‌های هر گروه و σ بیانگر انحراف استاندارد آن‌ها در همان گروه می‌باشد. در مرحله‌ی اول، ابتدا انتخاب ویژگی از داده‌های میکروآرایه انجام گردید و بر اساس ژن‌های انتخاب شده (۵۰ تا ۲۰۰ ژن دارای بالاترین رتبه)، شبکه‌ی ژن ساخته شد. بعد از ساخت شبکه‌ی ژن، ویژگی‌های توپولوژیک آن استخراج شد (جدول ۱) و ماتریس‌هایی که حاوی مقادیر کمی از هر ویژگی توپولوژیک هستند، به دست آمد.

در مرحله‌ی دوم، شبکه‌ی ژن از ماتریس داده‌ی ویژگی توپولوژیک با بیشترین تعداد ژن بازسازی شد که به دنبال آن ماتریس‌های ویژگی توپولوژیک بعد از انتخاب ویژگی از ماتریس داده‌ی ویژگی توپولوژیک با بیشترین تعداد ژن ایجاد گردید. در مطالعه‌ی حاضر به دلیل این‌که بیشترین تعداد ژن برای ماتریس داده‌ی توپولوژیک استخراج شد، بعد از ساخت شبکه‌ی ژن از داده‌های میکروآرایه، ماتریس حاوی ۲۰۰ ژن بود. بنابراین، در مرحله‌ی دوم

جدول ۱. ویژگی‌های توپولوژیک متداول در توصیف گراف‌ها (۲۵-۲۴)

ویژگی	تعریف
درجه‌ی تمرکز	معرف تعداد اتصالات یک گره از نظر مرکزیت بودن گره نسبت به گره‌های متصل به آن است.
توزیع درجه	تعداد اتصالات یک گره در شبکه‌ی ژن با نرمالیزه کردن اتصالات مربوط به مرکزیت یا تمرکز ژن از نظر اهمیت اتصال
بینابین بودن	معیاری است که نشان می‌دهد چه تعداد مسیر کوتاه از طریق یک گره عبور می‌کند و تعیین می‌نماید که کدام گره اثر واسطه‌ای بیشتری روی گره‌های دیگر دارد.
نزدیکی گره	عکس دوری تعریف می‌شود که به مجموع فواصل تمام گره‌های دیگر به یک ژن را بیان می‌کند.
ضریب خوشه‌بندی	معیاری است برای بیان درجه‌ای که با آن گره‌ها در یک گراف تمایل دارند در یک خوشه کنار هم باشند.
ویژه بودن گره	بردار ویژه‌ی ژن‌ها معیاری از تأثیر و نقش یک گره در یک شبکه‌ی ژن است.



شکل ۲. قانون عملکرد SVM (Support vector machines) (۲۶)

انتخاب ویژگی از روی ماتریس، داده‌ی توپوژلوژی با ۲۰۰ ژن انجام گرفت. بر خلاف مرحله‌ی قبل که به طور مستقیم انتخاب ویژگی روی داده‌ی بیان ژن اعمال شد، در این مرحله انتخاب ژن‌های شاخص بر اساس ویژگی‌های توپولوژیک انجام گرفت. با اعمال الگوریتم انتخاب ویژگی بر روی داده‌های ویژگی توپولوژیک، ژن‌های مورد نظر برای ساخت شبکه‌ی ژن تعیین شدند. بنابراین، شبکه‌ی ژن جدید بر اساس ژن‌های انتخاب شده از داده‌های ویژگی توپولوژیک ساخته شد.

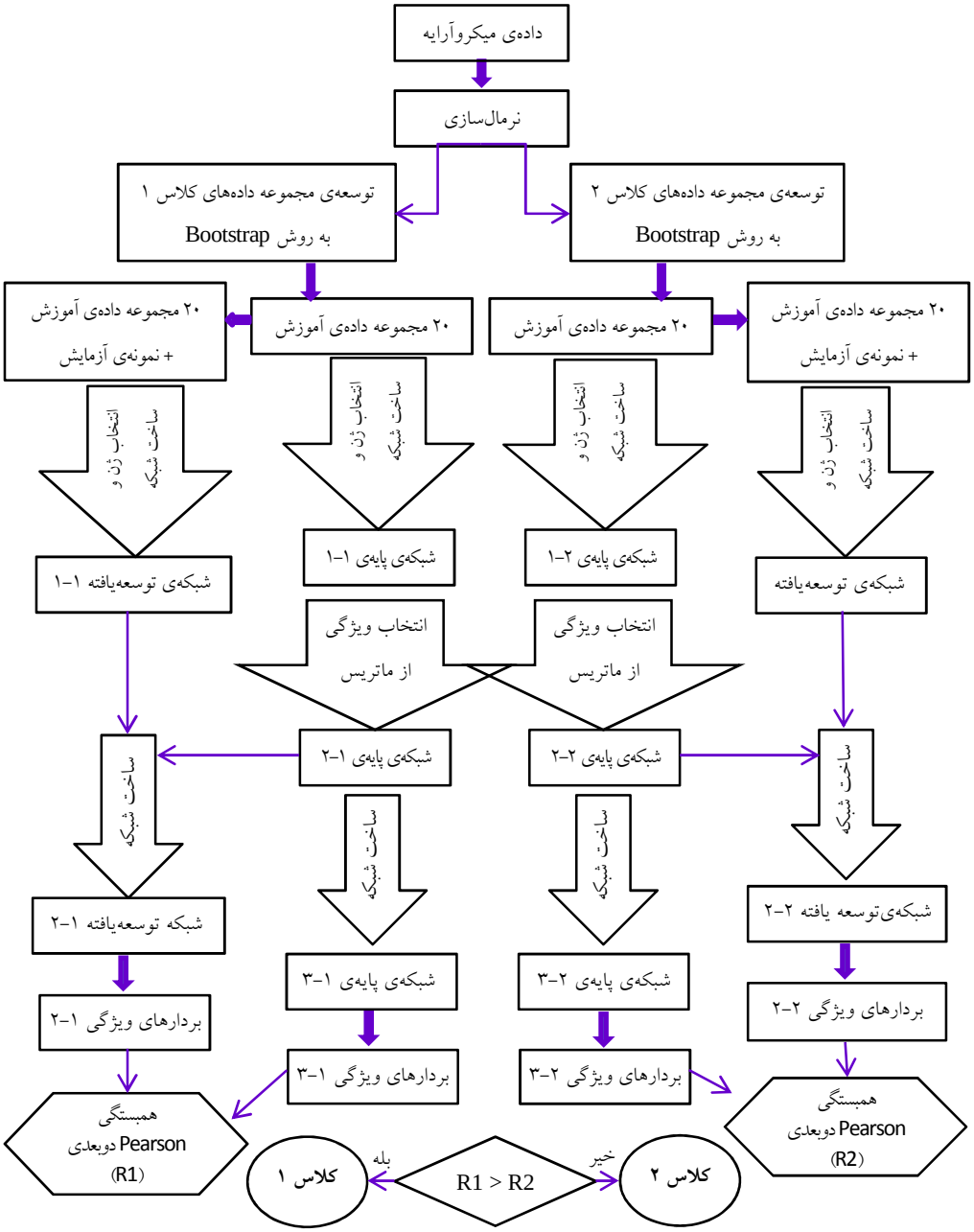
با توجه به این‌که ژن‌های شاخص انتخاب شده در مرحله‌ی دوم از شبکه شامل ۲۰۰ ژن بود، نمی‌توان برای مرحله‌ی آزمایش فقط از این ژن‌ها برای ساخت شبکه استفاده کرد. بنابراین، اگر به گونه‌ای از ژن‌های متصل به ژن‌های شاخص انتخاب شده به طور اولیه در طبقه‌بندی استفاده شود، به نظر می‌رسد قدرت طبقه‌بندی کننده در طبقه‌بندی نمونه‌های سرطانی بهبود یابد. برای دستیابی به این هدف کافی است که تمام ژن‌های متصل به ژن‌های اولیه‌ی انتخاب شده، برای ساخت شبکه‌ی جدید به کار گرفته شوند. معیار اتصال ژن‌ها در این مورد، وجود همبستگی بیش از یک آستانه‌ی مشخص بین پروفایل داده‌ی ویژگی توپولوژیک می‌باشد. لازم به ذکر است که ماتریس‌های ویژگی توپولوژیک شامل داده‌های توپولوژیک از ۲۰ فایل ۱۰۰ نمونه‌ای هستند. در پژوهش حاضر به دلیل این‌که شبکه‌ی ژن برای دو گروه سرطان به طور جداگانه ساخته شد، برای هر دو گروه کم‌خطر و پرخطر، ۲۰ بردار ویژگی توپولوژیک وجود داشت که ستون‌های ماتریس بردارهای ویژگی توپولوژیک از هر ۲۰ فایل ۱۰۰ نمونه‌ای تشکیل شده بود. در صورتی که ستون‌های ماتریس در ماتریس داده‌ی بیان ژن را نمونه‌های سرطانی تشکیل می‌دهند، اما در ماتریس داده‌ی توپولوژیک، هر ستون ماتریس در واقع یک شبکه‌ی ژن ساخته شده از یک مجموعه‌ی ۱۰۰ نمونه‌ای می‌باشد؛ چرا که از یک نمونه نمی‌توان در ساخت شبکه‌ی ژن استفاده نمود، بلکه به مجموعه‌ای از نمونه‌ها نیاز است. به عبارت دیگر، یک ویژگی توپولوژیک در شبکه‌ی ژن، برای یک نمونه تعریف نمی‌شود، بلکه به مجموعه‌ای از نمونه‌ها نسبت داده می‌شود.

طبقه‌بندی نمونه‌ها: در مطالعه‌ی حاضر به منظور تفکیک دو گروه کم‌خطر و پرخطر در بین نمونه‌ها، از ویژگی‌های توپولوژیک شبکه به جای مقدار بیان ژن استفاده گردید. این امر با توجه به توضیحات بخش قبل نیازمند رویکرد خاصی برای طبقه‌بندی می‌باشد که در شکل ۲ آمده است. مطابق شکل ۳، برای طبقه‌بندی نمونه‌های آزمایش، ابتدا داده‌های بیان ژن از یک نمونه‌ی آزمایش به کل داده‌های آموزش در هر دو گروه اضافه و شبکه‌های ژن متناظر بازسازی شد. شبکه‌ی ژن بازسازی شده

متشکل از داده‌های یک نمونه‌ی آزمایش و سایر داده‌های آموزشی مربوط به یکی از دو گروه نمونه می‌باشد.

پس از ساخت شبکه از داده‌های آموزش (شبکه‌ی پایه) و شبکه‌ی ژن بازسازی شده و بعد از اضافه شدن نمونه‌ی آزمایش به داده‌ی آموزش (شبکه‌ی توسعه یافته)، داده‌های توپولوژیک شبکه‌ی ژن ساخته شده از نمونه‌ی آزمایش به داده‌های ویژگی توپولوژیک اضافه گردید و شبکه‌ی ژن توسعه یافته‌ی جدیدی بازسازی شد.

همچنین، از داده‌های توپولوژیک شبکه‌ی ژن پایه هم دوباره شبکه‌ی ژن جدیدی بازسازی گردید. در مرحله‌ی بعد، ضرایب همبستگی دو بعدی بین بردارهای ویژگی توپولوژیک شبکه‌های ژن پایه (ساخته شده از شبکه‌ی ژن پایه) و بازسازی شده (شبکه‌ی بازسازی شده بعد از اضافه شدن داده‌های توپولوژیک شبکه‌ی ژن ساخته شده از نمونه‌ی آزمایش به داده‌های ویژگی توپولوژیک شبکه‌ی پایه) محاسبه شد.



شکل ۳. چارچوب کار طبقه‌بندی نمونه‌های آزمایش بر اساس ساخت شبکه‌ی ژن و ضریب همبستگی بین شبکه‌ها

برای هر نمونه‌ی آزمایشی اضافه شده، ضریب همبستگی بین دو شبکه‌ی ژن حاصل از افزودن نمونه به یکی از گروه‌ها محاسبه شد. سپس بر اساس ضرایب همبستگی به دست آمده برای تمام نمونه‌های آزمایش بین دو گروه، در مورد برچسب نمونه تصمیم‌گیری گردید. در این خصوص نمونه‌ای که ضریب همبستگی محاسبه شده‌ی بین شبکه‌های آن در نتیجه‌ی اضافه شدن به مجموعه‌ی داده‌های آموزشی یک گروه سرطان بیشتر از گروه دیگر بود، جزء همان گروه در نظر گرفته می‌شد. در مطالعه‌ی حاضر برای سنجش یافته‌های حاصل از روش پیشنهادی، از طبقه‌بندی کننده‌های غیر خطی k-NN (k-nearest neighbor) (۲۷) و SVMs (Support vector machines) (۲۸) استفاده شد که در ادامه به معرفی مختصر این دو طبقه‌بندی کننده پرداخته شد.

طبقه‌بندی کننده‌ی k-NN: یکی از بهترین طبقه‌بندی‌ها، طبقه‌بندی کننده‌ی k-NN است (۲۶). این طبقه‌بندی نمونه‌ی تست را متعلق به کلاسی می‌داند که بیشترین آرا را در بین k نزدیک‌ترین همسایگان آن داشته باشد. برای به دست آوردن نزدیک‌ترین همسایگان یک نمونه، اغلب از فاصله‌ی اقلیدسی طبق رابطه‌ی ۲ استفاده می‌شود.

$$deucl\,(x,\,t) = \sqrt{\sum_{i=1}^{m} d^{i}_{eucl}(x,\,t)} \qquad \text{رابطه‌ی ۲}$$

اگر مقادیر خصوصیات عددی و پیوسته باشد، Deuclid از رابطه‌ی ۳ به دست می‌آید.

$$deucl\,(x,\,t) = (ai\,(x) - ai\,(t))^{2} \qquad \text{رابطه‌ی ۳}$$

طبقه‌بندی کننده‌ی k-NN به دلیل قابلیت درک بالا و عدم نیاز به ایجاد فرضیه روی داده‌ها، روش ساده و پرکاربردی محسوب می‌شود. در این مطالعه طبقه‌بندی با تعداد همسایگی ۳ در داده‌ی بیان ژن و همسایگی ۱ در داده‌ی توپولوژی انجام گرفت. تعداد همسایگی ۱ برای داده‌ی توپولوژی به دلیل محدودیت تعداد ستون‌های داده‌ی توپولوژی بود.

طبقه‌بندی کننده‌ی SVMs (۲۹): طبقه‌بندی با SVM می‌تواند در مواقعی که داده‌ها به دقت به صورت دو کلاسی هستند، استفاده شود. SVM داده‌ها را با یافتن مهم‌ترین ابرصفحه که تمام نقاط داده از یک کلاس را از کلاس دیگر جدا می‌کند، طبقه‌بندی می‌نماید. مهم‌ترین ابرصفحه برای SVM به مفهوم بزرگ‌ترین حاشیه بین دو کلاس داده می‌باشد. حاشیه به مفهوم فاصله یا عرض بین دو خط موازی با ابرصفحه‌ای است که هیچ نقطه‌ای روی آن قرار ندارد (شکل ۲). بردارهای پشتیبان به نقاطی گفته می‌شود که به ابرصفحه یا خط جداساز نزدیک هستند. این نقاط روی مرز خط جداساز قرار دارند.

رابطه‌ی ۴ یک ابرصفحه را تعریف می‌کند:

رابطه‌ی ۴ $\langle w, x \rangle + b = 0$ در این رابطه x یک نقطه روی مرز تصمیم‌گیری (ابرصفحه) و w یک بردار عمود بر مرز تصمیم‌گیری و b مقدار Bias و مقداری حقیقی است. همان‌گونه که در شکل ۱ نشان داده شده است، $b/\|w\|$ بیانگر فاصله‌ی مبدأ تا مرز تصمیم‌گیری و $\langle w, x \rangle$ بیانگر ضرب داخلی دو بردار w و x می‌باشد. از آن‌جا که با ضرب یک ضریب ثابت در هر دو طرف معادله‌ی فوق باز هم تساوی برقرار است، بنابراین برای تعریف یکتای مقدار b و w شرایط زیر بر روی آن‌ها اعمال می‌شود:

رابطه‌ی ۵ اگر xi یک بردار پشتیبان باشد. $y\,(wxi\,b) \geq 1$

اگر xi یک بردار پشتیبان نباشد. $y\,(wxi\,b) > 1$

برای افزایش حاشیه، باید مقدار w کمینه شود. طبق شرایط فوق می‌توان رابطه‌ی ۶ را نتیجه گرفت.

$$W^{T}x_i + b > 1 \qquad \text{رابطه‌ی ۶}$$

برای راحتی کار و استفاده از جبر خطی، به جای کمینه کردن w، تابع $1/2||w||^2$ کمینه می‌گردد. با یادآوری توابع لاگرانژ و شروط Karush-Kuhn-Tucker، اگر بخواهد تابعی مثل F(x) نسبت به متغیر مستقل خود یعنی x کمینه شود به شرطی که $g(x) \geq 0$ باشد، می‌توان از تابع لاگرانژ (رابطه‌ی ۷) استفاده نمود.

$$L(x, u) = F(x) - UG(x); \quad U \geq 0 \qquad \text{رابطه‌ی ۷}$$

که در این تابع باید x کمینه و U بیشینه شود. بنابراین، شروط Karush-Kuhn-Tucker به صورت روابط ۸ و ۹ خواهد بود.

$$\frac{\partial L(x,u)}{\partial (x)} = 0 \qquad \text{رابطه‌ی ۸}$$

$$UG(x) = 0 \qquad \text{رابطه‌ی ۹}$$

حال اگر تابع $1/2\|w\|^2$ را به جای F(x) فرض کنیم و این تابع را در روابط فوق (۶ و ۷) جایگذاری کنیم، روابط ۱۰ تا ۱۶ را داریم.

$$w = \sum_{i}^{m} a_i y_i x_i \qquad \text{رابطه‌ی ۱۰}$$

$$\sum_{i}^{m} a_i y_i = 0 \qquad \text{رابطه‌ی ۱۱}$$

$$Q(w, b, a) = \sum_{i}^{m} a_i [y_i (w x_i - b) - 1] = \frac{1}{2}||w||^2 \qquad \text{رابطه‌ی ۱۲}$$

$$Q(a) = 1/2 \sum_i \sum_j a_i a_j . y_i y_j x^T x + \sum_i a_i \qquad \text{رابطه‌ی ۱۳}$$

$$\text{Minimize } 1/2 \sum_i \sum_j a_i a_j . y_i y_j x^T x - \sum_i a_i \qquad \text{رابطه‌ی ۱۴}$$

$$H = y_i y_j x x_i^T x_j \qquad \text{رابطه‌ی ۱۵}$$

اگر جدایی‌پذیری به صورت خطی نباشد، نمونه‌ها به یک فضا با بعد بالا نگاشت داده می‌شود که در فضای جدید نمونه‌ها می‌توانند

به صورت خطی از هم جدا شوند. طبق رابطه‌ی ۱۵، برای نگاشت به یک تابع هسته یا Kernel نیاز است.

در نهایت، تابع نگاشت یا تبدیل فضای محدود به فضای دارای ابعاد بالا به صورت رابطه‌ی ۱۶ تعریف می‌شود.

$$G(x) = w^T\varphi(x) + b \qquad \text{رابطه‌ی ۱۶}$$

در مطالعه‌ی حاضر پارامتر خروجی آلفا همان ضریب لاگرانژ بود که در روابط نوشته شده است و مقدار Bias (b)، ۱/۸۳۳- بود.

برای ارزیابی صحت طبقه‌بندی در تعداد ژن‌های مختلف انتخاب شده، از روش روایی‌سنجی متقاطع 5-fold استفاده شد (۲۹)؛ بدین صورت که نمونه‌های سرطانی مورد مطالعه از هر دو گروه سرطانی کم‌خطر و پرخطر به پنج قسمت مساوی تقسیم شدند و در هر بار ارزیابی برای محاسبه‌ی صحت طبقه‌بندی، یک قسمت از داده‌های هر گروه به عنوان داده‌ی آزمایش و چهار قسمت باقی‌مانده به عنوان داده‌ی آموزش در نظر گرفته شد. با تکرار ۵ بار این فرایند و چرخش نمونه‌های آزمایش در هر تکرار، تمام نمونه‌ها یک بار به عنوان داده‌ی آزمایش مورد بررسی قرار گرفتند. بعد از محاسبه‌ی صحت طبقه‌بندی برای هر بخش از داده‌ها در ۵ بار تکرار طبقه‌بندی، مقدار میانگین صحت‌های محاسبه شده برای ۵ بار محاسبه شد و به همراه مقدار انحراف استاندارد صحت نهایی ارایه گردید.

یافته‌ها

برای ساخت شبکه‌ی ژن به شیوه‌ی مدل ارتباطی از جعبه‌ی ابزار فوگا (۳۰) در نرم‌افزار MATLAB استفاده شد. بدین منظور از ضریب همبستگی Pearson به عنوان آستانه‌ی تعیین یال بین گره‌ها استفاده گردید (۰/۸۰). با توجه به امتیاز اختصاص یافته به ژن‌ها در رتبه‌بندی انجام شده بر اساس ویژگی‌های توپولوژیک مختلف و به منظور انتخاب ژن‌های شاخص، ویژگی توپولوژیک بینابینی مورد استفاده قرار گرفت. از آن جایی که در این روش امتیازدهی توزیع درجه بعد از ویژگی بینابینی در رتبه‌ی دوم قرار داشت، در ادامه‌ی ساخت شبکه‌ی ژن بر اساس داده‌های ویژگی توپولوژیک، توزیع درجه نیز انجام شد. برای این کار، بر اساس رتبه‌بندی انجام گرفته مبتنی بر ویژگی توزیع درجه، ۲۰ ژن با بالاترین رتبه برای ساخت شبکه‌ی ژن انتخاب شد. جهت توسعه‌ی مجموعه ژن‌های اولیه، ژن‌هایی که با همبستگی Pearson بالای ۰/۳۵ به ۲۰ ژن اولیه متصل بودند، به مجموعه اولیه اضافه شدند. بنابراین، ساخت شبکه‌ی ژن با انتخاب ژن‌های متصل به ۲۰ ژن انتخاب شده از ماتریس ویژگی توپولوژیک، توزیع درجه ادامه یافت.

در مرحله‌ی اول ارزیابی نتایج، مشابه با روش Liu و همکاران (۱۷-۱۶)، ابتدا تعداد ژن‌های مستخرج از میکروآرایه توسط یک روش انتخاب ویژگی کاهش یافت. پس از ساخت شبکه با ژن‌های منتخب، از طریق آستانه‌گذاری بر ضریب همبستگی بین مقدار ویژگی توپولوژیک در شبکه‌های متناظر با دو گروه نمونه‌ها طبقه‌بندی انجام گرفت. در این مرحله صحت طبقه‌بندی نمونه‌ها بر اساس ارزیابی متقاطع 5-fold محاسبه گردید و میانگین و انحراف معیار آن‌ها در جدول ۲ ارایه شده است. ساخت شبکه با تعداد مختلف ژن (سطرهای جدول) انجام گرفت و در مرحله‌ی طبقه‌بندی از ویژگی‌های مختلف توپولوژی (ستون‌های جدول) به طور مستقل استفاده شد.

جدول ۲. صحت طبقه‌بندی بر اساس ویژگی‌های توپولوژیک شبکه درتعداد ژن‌های مختلف با انتخاب ویژگی از داده‌ی بیان ژن

تعداد	توزیع درجه	درجه‌ی تمرکز	بینابینی	ضریب خوشه‌بندی	نزدیکی رئوس	گره‌ی ویژه
۶۰	۶۳/۵ ± ۷	۶۵/۰ ± ۱۰	۷۳/۵ ± ۲	۸۲/۰ ± ۶	۴۸/۰ ± ۵	۷۴/۵ ± ۱۵
۷۰	۵۷/۰ ± ۵	۵۹/۰ ± ۱۶	۶۹/۰ ± ۱۰	۷۶/۰ ± ۱۲	۴۴/۰ ± ۴	۶۶/۰ ± ۸
۸۰	۷۳/۵ ± ۶	۸۱/۰ ± ۶	۵۹/۰ ± ۶	۴۱/۰ ± ۱۳	۸۶/۰ ± ۱۳	۸۸/۵ ± ۴
۹۰	۶۳/۵ ± ۱۵	۷۰/۰ ± ۱۵	۵۳/۰ ± ۱۱	۳۸/۰ ± ۲۰	۷۸/۵ ± ۱۰	۷۶/۰ ± ۱۰
۱۰۰	۵۹/۰ ± ۱۳	۷۴/۰ ± ۱۱	۵۸/۵ ± ۱۲	۵۷/۵ ± ۱۶	۶۹/۵ ± ۸	۶۴/۰ ± ۱۷
۱۲۰	۵۹/۴ ± ۹	۷۱/۰ ± ۷	۸۴/۰ ± ۵	۶۸/۰ ± ۲۱	۷۹/۰ ± ۹	۶۷/۰ ± ۹
۱۴۰	۶۴/۹ ± ۵	۷۳/۰ ± ۷	۸۲/۵ ± ۷	۴۹/۰ ± ۶	۷۸/۰ ± ۱۱	۷۴/۰ ± ۷
۱۶۰	۷۰/۵ ± ۱۶	۷۸/۵ ± ۹	۷۷/۰ ± ۳	۶۱/۰ ± ۷	۸۲/۰ ± ۹	۷۴/۵ ± ۱۷
۲۰۰	۷۳/۲ ± ۶	۸۰/۰ ± ۸	۶۴/۵ ± ۴	۶۱/۰ ± ۵	۷۹/۵ ± ۵	۶۶/۰ ± ۱۱

جدول ۳. صحت طبقه‌بندی براساس ویژگی‌های توپولوژیک شبکه‌ی ژن در تعداد ژن‌های مختلف انتخاب شده با انتخاب ویژگی از داده‌ی ویژگی توپولوژیک بینابینی

تعداد	توزیع درجه	درجه‌ی تمرکز	بینابینی	ضریب خوشه‌بندی	نزدیکی رئوس	گره‌ی ویژه
۲۰	۹۲/۰ ± ۱	۹۲/۵ ± ۱	۲۲/۰ ± ۳۷	۹۱/۵ ± ۱	۹۲/۰ ± ۱	۹۱/۰ ± ۵
۳۰	۸۷/۵ ± ۱۲	۸۷/۵ ± ۱۲	۱۹/۰ ± ۳۸	۸۷/۰ ± ۱۲	۸۶/۰ ± ۱۳	۹۲/۴ ± ۸
۴۰	۹۳/۰ ± ۱	۹۳/۰ ± ۱	۱۸/۵ ± ۳۸	۹۲/۵ ± ۱	۹۲/۰ ± ۱	۹۴/۰ ± ۱
۵۰	۸۸/۰ ± ۱۲	۸۸/۰ ± ۱۲	۱/۵ ± ۱	۸۷/۰ ± ۱۲	۸۶/۰ ± ۱۳	۹۰/۰ ± ۱۳
۶۰	۷۲/۰ ± ۴۲	۶۵/۰ ± ۴۵	۲۲/۰ ± ۴۳	۸۶/۹ ± ۱۲	۶۴/۰ ± ۴۴	۶۵/۵ ± ۴۵
۷۰	۹۳/۵ ± ۱	۹۳/۰ ± ۱	۱/۱ ± ./۵	۸۸/۰ ± ۱۲	۹۲/۰ ± ۱	۹۴/۰ ± ۲

جدول ۴. صحت طبقه‌بندی براساس ویژگی‌های توپولوژیک شبکه‌ی ژن درتعداد ژن‌های مختلف انتخاب شده با انتخاب ژن از داده‌ی ویژگی توپولوژیک توزیع درجه و توسعه‌ی مجموعه‌ی ژن‌ها بر اساس همبستگی ژن‌ها از روی داده‌های درجه‌ی اتصال

تعداد	توزیع درجه	درجه‌ی تمرکز	بینابینی	ضریب خوشه‌بندی	نزدیکی رئوس	گره‌ی ویژه
۲۰	۷۳ ± ۲۰	۷۴ ± ۱۹	۱۷/۵ ± ۲۰	۸۷ ± ۱۰	۷۳ ± ۲۰	۷۳ ± ۱۹
۷۱	۹۰ ± ۱۳	۹۰ ± ۱۳	۲/۰ ± ۱	۸۸ ± ۱۲	۸۸ ± ۱۲	۹۱ ± ۱۲

در مرحله‌ی دوم ارزیابی که نتایج آن در جدول ۳ گزارش شده است، برای ساخت شبکه‌ی ژن از ۲۰۰ ژن داده‌ی ویژگی توپولوژیک استفاده گردید و در مرحله‌ی طبقه‌بندی تنها از ژن‌های انتخاب شده بر اساس ویژگی‌های توپولوژیک بهره گرفته شد. لازم به ذکر است که در این مرحله نیز طبقه‌بندی بر اساس آستانه‌گذاری همبستگی بین ویژگی‌های توپولوژی در دو شبکه انجام گرفت. تنها تفاوت در روش حصول جداول ۲ و ۳ که اختلاف زیادی از نظر نتایج دارند، جابه‌جایی مرحله‌ی انتخاب ویژگی از مرحله‌ی قبل از ساخت شبکه به پس از ساخت شبکه است. به عبارت دیگر، رتبه بندی آن‌ها در جدول ۲ بر اساس مقدار بیان ژن و در جدول ۳ بر اساس ویژگی‌های توپولوژی انجام گرفته است.

ارزیابی دیگری به منظور بررسی تأثیر انتخاب ویژگی مبتنی بر ویژگی‌های توپولوژی انجام گرفت. در این ارزیابی که نتایج آن در جدول ۴ گزارش شده است، ابتدا ۲۰ ژن از شبکه‌ی ساخته شده از کل ژن‌ها بر اساس رتبه‌بندی ویژگی توزیع درجه انتخاب شد و سپس، ژن‌های متصل به این ۲۰ ژن (اتصال به معنی وجود ضریب همبستگی بالای ۰/۳۵ بین دو ژن) مطابق آنچه در ابتدای بخش شرح داده شد، انتخاب و بار دیگر شبکه‌ای با استفاده از مجموعه ژن‌های توسعه یافته (۷۱ ژن) از ابتدا ساخته شد. در انتها، با مقایسه‌ی همبستگی بین ویژگی‌های توپولوژی مختلف در شبکه‌های بازسازی شده با استفاده از مجموعه‌ی توسعه یافته، مقدار صحت طبقه‌بندی محاسبه گردید و مطابق جدول ۴ گزارش شد. به منظور بهبود بیشتر مقدار صحت طبقه‌بندی در مرحله‌ی بعدی ارزیابی، به جای آستانه‌گذاری خطی از طبقه‌بندی کننده‌های غیر خطی k-NN و ماشین بردار پشتیبان با هسته (Kernel) چند جمله‌ای استفاده شد. مطابق نتایج جداول ۳ و ۴، بیشترین میانگین صحت طبقه‌بندی توسط ویژگی، توزیع درجه و ویژه بودن گره (ژن) به دست آمده است.

در جداول ۵ و ۶ نتایج طبقه‌بندی k-NN با اختلاف اندک بین مقادیر صحت طبقه‌بندی با ویژگی درجه‌ی تمرکز ژن و ویژه بودن گره، تنها برای ویژگی درجه‌ی تمرکز ژن در نظر گرفته شد. در جداول ۵ و ۶، میانگین و انحراف معیار صحت طبقه‌بندی حاصل از ارزیابی متقاطع با اعمال طبقه‌بندی کننده‌ی k-NN و SVM برای حالتی که انتخاب ویژگی قبل از بازسازی شبکه انجام شده، نشان داده شده است. در جداول ۷ و ۸ نتایج مشابه برای حالتی که انتخاب ویژگی پس از بازسازی شبکه بر ویژگی‌های توپولوژی انجام شد، گزارش گردید.

مطابق جداول ۳ و ۴، ژن‌های شاخص عود سرطان سینه که بر اساس ویژگی‌های توپولوژیک انتخاب شدند، صحت پیش‌گویی مناسب‌تری نسبت به شاخص‌های مبتنی بر بیان ژن دارند. اکنون باید ژن‌های انتخابی از نظر آنتولوژی نیز تحلیل گردد تا نقش آن‌ها در بروز پدیده‌ی عود بهتر مشخص شود. ارزیابی آنتولوژی ژن‌ها در سه سطح جزء سلولی، فرایند زیستی و عملکرد در سطح مولکولی توسط نرم‌افزار EASE نسخه‌ی ۲ (۳۱) انجام گرفت. نتایج حاصل از تحلیل آنتولوژی بر روی ژن‌های به دست آمده با اعمال انتخاب ویژگی قبل از ساخت شبکه، انتخاب با معیار ویژگی توپولوژی بینابینی و ویژگی توزیع درجه به ترتیب در شکل‌های ۶-۴ نمایش داده شده است. در این شکل‌ها تعداد ژن‌های شاخصی که در هر یک از شاخه‌های آنتولوژی مختلف قرار گرفته است، مشخص گردید.

جدول ۵. صحت طبقه‌بندی k-NN (k-nearest neighbor) در تعداد ژن‌های مختلف انتخاب شده با انتخاب ژن از داده‌ی ویژگی توپولوژیک بینابینی

تعداد	توزیع درجه	درجه‌ی تمرکز	بینابینی	ضریب خوشه‌بندی	نزدیکی رئوس	گره‌ی ویژه
۲۰	۹۹/۹ ± ۰/۰۵	۹۹/۹ ± ۰/۰۵	۵۵/۰ ± ۲۴	۹۹/۹ ± ۰/۰۵	۹۹/۹ ± ۰/۰۵	۹۹/۹ ± ۰/۰۵
۳۰	۱۰۰	۱۰۰	۵۱/۵ ± ۲۷	۱۰۰	۱۰۰	۱۰۰
۴۰	۱۰۰	۱۰۰	۵۵/۰ ± ۲۴	۱۰۰	۱۰۰	۱۰۰
۵۰	۱۰۰	۱۰۰	۴۰/۵ ± ۸	۱۰۰	۱۰۰	۱۰۰
۶۰	۹۱/۰ ± ۱۹	۹۱ ± ۱۹	۴۰/۵ ± ۸	۹۱/۰ ± ۱۹	۸۹/۰ ± ۲۵	۸۶/۰ ± ۲۷
۷۰	۱۰۰	۱۰۰	۴۴/۴	۱۰۰	۱۰۰	۱۰۰

جدول ۶. صحت طبقه‌بندی SVM (Support vector machines) در تعداد ژن‌های مختلف انتخاب شده با انتخاب ژن از داده‌ی ویژگی توپولوژیک بینابینی

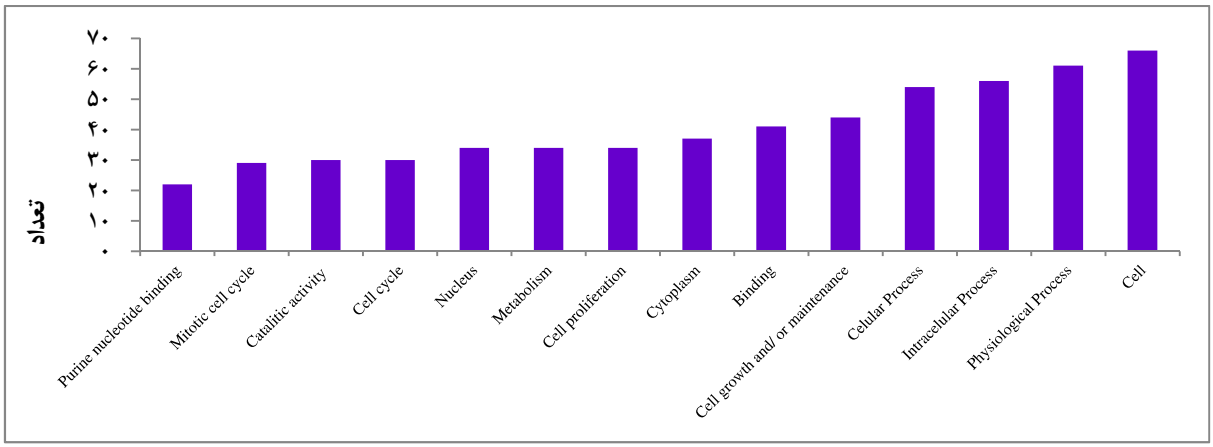
تعداد	توزیع درجه	درجه‌ی تمرکز	بینابینی	ضریب دی	نزدیکی رئوس	گره‌ی ویژه
۲۰	۹۹/۹۵ ± ۰/۰۵	۱۰۰	۵۴/۰ ± ۲۲	۱۰۰	۹۹/۹ ± ۰/۱	۹۹/۹ ± ۰/۱۰
۳۰	۱۰۰	۹۸/۹ ± ۱/۵	۵۱/۵ ± ۲۷	۱۰۰	۱۰۰	۹۹/۹۸ ± ۰/۰۴
۴۰	۹۹/۹۵ ± ۰/۰۵	۹۸/۹ ± ۱/۵	۵۵/۰ ± ۲۴	۱۰۰	۱۰۰	۹۹/۹ ± ۰/۲۵
۵۰	۱۰۰	۹۸/۹ ± ۱/۷	۴۰/۵ ± ۸	۱۰۰	۱۰۰	۹۹/۸ ± ۰/۳۰
۶۰	۸۹/۰ ± ۲۱/۰۰	۸۴/۵ ± ۱۹/۰	۳۱/۰ ± ۳۱	۸۹ ± ۲۱	۸۶/۰ ± ۲۷/۰	۸۶/۰ ± ۲۷/۰۰
۷۰	۹۹/۹ ± ۰/۱۰	۹۷/۹ ± ۲/۵	۴۴/۴	۱۰۰	۱۰۰	۹۹/۶ ± ۰/۷۰

جدول ۷. صحت طبقه‌بندی توسط k-NN (k-nearest neighbor) در تعداد ژن‌های مختلف انتخاب شده با انتخاب ژن از داده‌ی بیان ژن

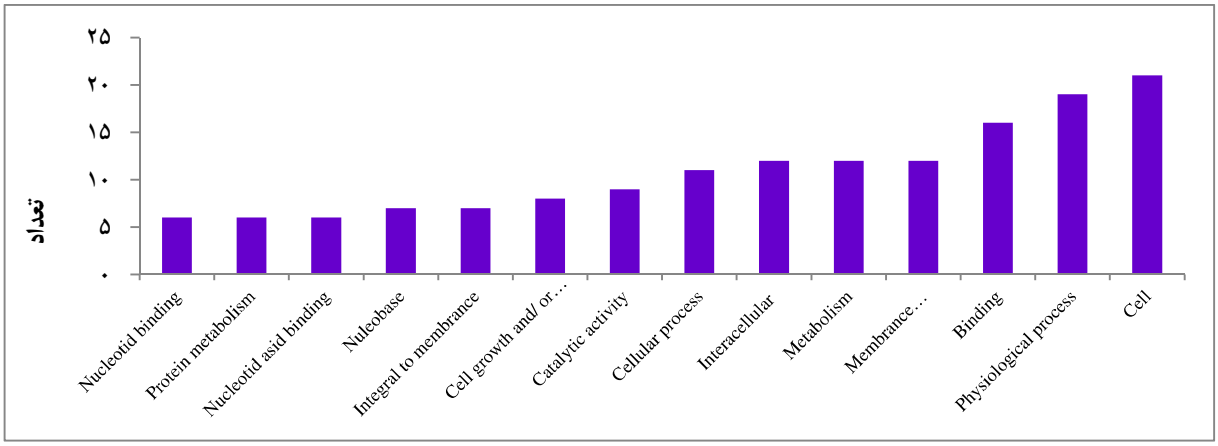
تعداد ژن	صحت
۶۰	۷۰ ± ۳
۷۰	۷۰ ± ۱
۸۰	۷۰ ± ۳
۹۰	۶۷ ± ۵
۱۰۰	۷۱ ± ۱
۱۲۰	۶۹ ± ۳
۱۴۰	۷۱ ± ۲
۱۶۰	۷۵ ± ۴
۲۰۰	۷۲/۵ ± ۵

جدول ۸ صحت طبقه‌بندی توسط SVM (Support vector machines) در تعداد ژن‌های مختلف انتخاب شده با انتخاب ژن از داده‌ی بیان ژن

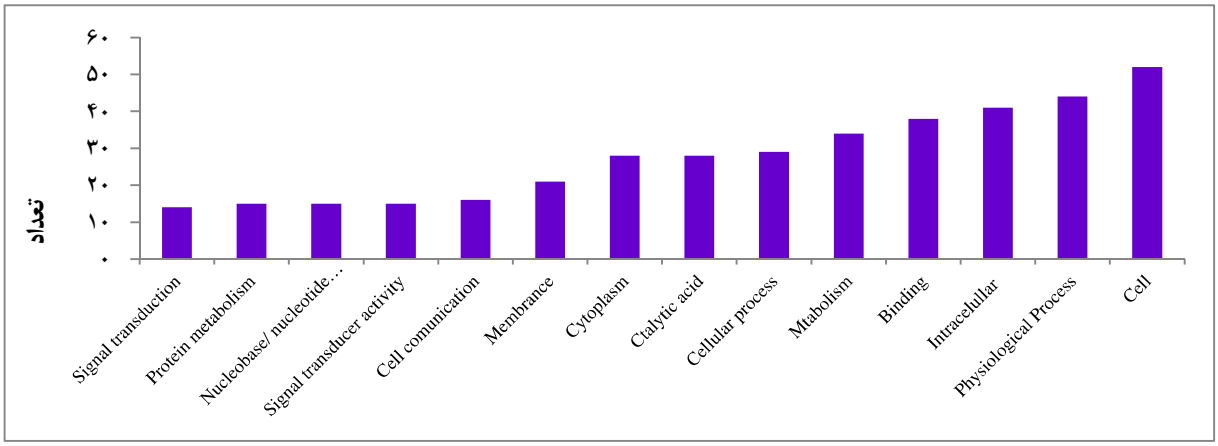
تعداد ژن	صحت
۶۰	۶۴/۵ ± ۲
۷۰	۶۶/۰ ± ۱
۸۰	۶۷/۰ ± ۳
۹۰	۶۸/۰ ± ۲
۱۰۰	۶۷/۵ ± ۱
۱۲۰	۶۷/۰ ± ۳
۱۴۰	۶۷/۰ ± ۲
۱۶۰	۶۹/۰ ± ۴
۲۰۰	۶۴/۰ ± ۵



شکل ۴. هستی‌شناسی ژن‌ها بر اساس انتخاب ویژگی از داده‌ی بیان ژن



شکل ۵. هستی‌شناسی ژن‌ها بر اساس انتخاب ویژگی از داده‌ی ویژگی توپولوژیک بینابینی



شکل ۶. هستی‌شناسی ژن‌ها بر اساس انتخاب ویژگی از داده‌ی ویژگی توپولوژیک درجه و توسعه‌ی مجموعه ژن‌ها بر اساس اطلاعات درجه‌ی اتصال ژن‌ها

بحث

در مطالعه‌ی حاضر، استفاده از ویژگی‌های توپولوژی در جهت رتبه‌بندی ژن‌ها و مشارکت آن‌ها در مدل طبقه‌بندی کننده بر اساس همبستگی بین شبکه‌ها مورد بررسی قرار گرفت. مقایسه‌ی نتایج جداول ۲ و ۳ نشان داد که ویژگی‌های توپولوژیک معیار مناسبی برای انتخاب ژن است و در جهت افزایش صحت طبقه‌بندی نمونه‌ها و کاهش بعد قبل از ساخت شبکه به منظور دستیابی به نتایج پایدار و کاهش هزینه‌ی محاسباتی مؤثر می‌باشد، اما اگر حذف ژن‌ها با معیار مناسبی انجام نگیرد، منجر به تغییر ویژگی‌های توپولوژیک و مخدوش شدن نتایج نهایی خواهد شد. همچنین، با یک بررسی از انتها به ابتدا، استفاده از توزیع درجه و انتخاب ژن‌های متصل به ژن‌های شاخص به عنوان معیار مناسبی در کاهش بعد مسأله با کمترین اثر سوء بر نتایج معرفی شد. مقایسه‌ی یافته‌های جداول ۳ و ۴ کاهش صحت طبقه‌بندی را مشخص نمود. نتایج جداول ۵ و ۶ تأثیر مثبت کاربرد طبقه‌بندی کننده‌های غیر خطی (k-NN و SVM) را نسبت به یک روش خطی در تولید نتایج مبتنی بر انتخاب ویژگی از داده‌ی بیان ژن با افزایش خفیف در صحت طبقه‌بندی (جدول ۲) را نشان می‌دهد. جداول ۷ و ۸ نیز نشان دهنده‌ی مزیت کاربرد طبقه‌بندی کننده‌ی غیر خطی در کاربرد ویژگی‌های توپولوژیک برای طبقه‌بندی نسبت به طبقه‌بندی کننده‌ی خطی (جدول ۳) می‌باشد. مقایسه‌ی نتایج جداول ۵ و ۷ نیز تأیید دوباره‌ای بر مزیت کاربرد ویژگی‌های توپولوژی در افزایش صحت طبقه‌بندی نمونه‌ها است.

طبقه‌بندی نمونه‌های سرطانی بر اساس ویژگی‌های توپولوژیک شبکه‌ی ژن پیش‌تر در مطالعات Liu و همکاران (۱۷-۱۶) به طور مشابه از نظر بهره‌گیری از ساخت شبکه‌ی ژن و استخراج برخی ویژگی‌های توپولوژیک مورد بررسی قرار گرفته است. تفاوت مطالعه حاضر نسبت به مطالعات قبلی را می‌توان در مورد خاص مورد بررسی (پیش‌گویی عود سرطان سینه)، مرحله و معیار انتخاب ویژگی

و تعداد گروه‌های نمونه‌ها عنوان کرد. در مطالعه‌ی Liu و همکاران، انتخاب ژن و کاهش بعد قبل از ساخت شبکه‌ی ژن و از روی داده‌ی بیان ژن صورت گرفته بود (۱۷-۱۶)، اما در مطالعه‌ی حاضر انتخاب ژن‌های شاخص پس از ساخت شبکه‌ی ژن و از روی ویژگی‌های توپولوژیک انجام شد.

میانگین صحت طبقه‌بندی به دست آمده در مطالعه‌ی حاضر پس از کاربرد k-NN (۹۸/۵ درصد) و SVM (۹۶/۵ درصد) نسبت به مقدار حاصل از روش Liu و همکاران (۱۷-۱۶) در داده‌های مشابه، از داده‌های بیان ژن نمونه‌های سرطانی به ترتیب برای طبقه‌بندی کننده‌های k-NN (۷۰/۶ درصد) و SVM (۶۶/۷ درصد) بیشتر بود. در پژوهش Yang و همکاران از دو مجموعه داده‌ی سرطان سینه‌ی گروه مورد و شاهد، شبکه‌های ژنی بازسازی شدند که بعد از آن با تفاضل دو شبکه‌ی ژن، شبکه‌ی تفاضلی تشکیل شده بود (۱۹). آن‌ها از روی شبکه‌ی تفاضلی ایجاد شده، ژن‌ها را بر اساس مقدار ویژگی درجه (ژن‌های با درجه‌ی اتصال بزرگ‌تر از ۵) و P رتبه‌بندی کردند (۱۹) که مانند مطالعه‌ی حاضر، ویژگی توزیع درجه را یکی از معیارهای انتخاب ژن در نظر گرفتند (۱۹). در مطالعه‌ی Ahn و همکاران برای طبقه‌بندی نمونه‌ی سرطانی از ساخت شبکه‌ی ژن استفاده گردید. به این صورت که آن‌ها از ژن‌های متناظر با شبکه‌های پروتئین PPI (Protein-Protein interaction) و برهم کنش ژنتیک (Genetic interactions یا GI) و آنالیز مسیرهای بیولوژیک در ساخت شبکه‌ی ژن استفاده نمودند (۲۱)، اما همان‌گونه که در ابتدا اشاره شد، شبکه‌های استاندارد پروتئینی با وجود ارایه‌ی اطلاعات ارزشمند، داده‌های کاملی نیستند و عاری از خطا نمی‌باشند. بنابراین، ساخت شبکه‌ی ژن از روی پروفایل بیان ژن در میکروآرایه، می‌تواند نتایج مطمئن‌تری ارایه دهد.

بر اساس ارزیابی و تحلیل آنتولوژی ژن‌های شاخص به دست آمده از روش پیشنهاد شده در مطالعه‌ی حاضر (شکل‌های ۴-۲)، می‌توان در مورد نقش کارکردی ژن‌های شاخص و انطباق آن‌ها با کارکردهای دخیل در سرطان اظهار نظر کرد. از این نظر، با مبنا قرار دادن ویژگی توپولوژی بینابینی نسبت به ویژگی توزیع درجه‌ی انطباق بیشتر با کارکرد سلول‌های سرطانی شناخته شده، مشاهده می‌شود که این انطباق از نظر کلی مشابه حالتی می‌باشد که انتخاب ویژگی به طور مستقیم قبل از ساخت شبکه انجام گرفته است. نتیجه این تحلیل قابلیت تفسیر بیولوژی ژن‌های استخراج شده به عنوان شاخص پیش‌گویی را بیان می‌کند.

هدف بعدی در راستای ادامه این مطالعه، بررسی و مقایسه‌ی نتایج صحت طبقه‌بندی و تحلیل آنتولوژی آن‌ها با بهره گرفتن از سایر روش‌های ساخت شبکه است. همچنین، در ادامه‌ی تحقیق حاضر و با توجه به تحلیل ضرایب همبستگی بین شبکه‌های ژن متناظر با دو گروه مورد مطالعه، پیشنهاد می‌شود روش مناسب‌تری به جز تحلیل ضرایب همبستگی به منظور طبقه‌بندی یک نمونه آزمایش ارایه شود.

تشکر و قدردانی

مقاله‌ی حاضر برگرفته از پروژه‌ی پایان‌نامه کارشناسی ارشد مصوب دانشگاه علوم پزشکی اصفهان به شماره‌ی ۳۹۴۰۸۶ می‌باشد. بدین وسیله از معاونت تحقیقات و فن‌آوری دانشکده‌ی فن‌آوری‌های نوین علوم پزشکی به جهت تأمین هزینه‌ی اجرای پروژه قدردانی می‌گردد.

References

1. Tahergorabi Z, Moodi M, Mesbahzadeh B. Breast cancer: A preventable disease. J Birjand Univ Med Sci 2014; 21(2): 126-41. [In Persian].
2. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res 2005; 65(16): 7065-70.
3. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415(6871): 530-6.
4. Brambilla C, Fievet F, Jeanmart M, de Fraipont F, Lantuejoul S, Frappat V, et al. Early detection of lung cancer: role of biomarkers. Eur Respir J Suppl 2003; 39: 36s-44s.
5. Brennan DJ, O'Brien SL, Fagan A, Culhane AC, Higgins DG, Duffy MJ, et al. Application of DNA microarray technology in determining breast cancer prognosis and therapeutic response. Expert Opin Biol Ther 2005; 5(8): 1069-83.
6. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP. Classification of microarray data using gene networks. BMC Bioinformatics 2007; 8: 35.
7. Brazhnik P, de la Fuente A, Mendes P. Gene networks: how to put the function in genomics. Trends Biotechnol 2002; 20(11): 467-72.
8. Curtis RE, Yuen A, Song L, Goyal A, Xing EP. TVNViewer: an interactive visualization tool for exploring networks that change over time or space. Bioinformatics 2011; 27(13): 1880-1.
9. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000; 418-29.
10. Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput 1999; 17-28.
11. Gevaert O, De SF, Timmerman D, Moreau Y, De MB. Predicting the prognosis of breast cancer by integrating

clinical and microarray data with Bayesian networks. Bioinformatics 2006; 22(14): e184-e190.
12. Sakamoto E, Iba H. Inferring a system of differential equations for a gene regulatory network by using genetic programming. Proceedings of the 2001 Congress on Evolutionary Computation; 2001 May 27-30; Seoul, South Korea.
13. Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, Charnock-Jones DS, et al. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. Bioinformatics 2008; 24(7): 932-42.
14. Vohradsky J. Neural network model of gene expression. FASEB J 2001; 15(3): 846-54.
15. Segal E, Shapira M, Regev A, Peer D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003; 34(2): 166-76.
16. Liu CC, Chen WSE, Chang PC, Chen JJW. Topological-based classification using artificial gene networks. Proceedings of 4th IEEE Conference on Cognitive Informatics; 2005 Aug 8-10; Irvine, USA.
17. Liu CC, Chen WS, Lin CC, Liu HC, Chen HY, Yang PC, et al. Topology-based cancer classification and related pathway mining using microarray data. Nucleic Acids Res 2006; 34(14): 4069-80.
18. Raza K, Jaiswal R. Reconstruction and analysis of cancer-specific generegulatory networks from gene expression profiles. International Journal on Bioinformatics and Biosciences 2013; 3(2): 25-34.
19. Yang B, Zhang J, Yin Y, Zhang Y. Network-based inference framework for identifying cancer genes from gene expression data. Biomed Res Int 2013; 2013: 401649.
20. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007; 3: 140.
21. Ahn J, Yoon Y, Park C, Shin E, Park S. Integrative gene network construction for predicting a set of complementary prostate cancer genes. Bioinformatics 2011; 27(13): 1846-53.
22. Bockmayr M, Klauschen F, Gyorffy B, Denkert C, Budczies J. New network topology approaches reveal differential correlation patterns in breast cancer. BMC Syst Biol 2013; 7: 78.
23. Muszynski M, Osowski S. Data mining methods for gene selection on the basis of gene expression arrays. Int J Appl Math Comput Sci 2014; 24(3): 657-68.
24. Wang Y, Yao M, Yang J. NIM: a node influence based method for cancer classification. Comput Math Methods Med 2014; 2014: 826373.
25. Teng CY, Lin YR, Adamic L. Recipe recommendation using ingredient networks. Proceedings of Web Science; 2012 Jun 22-24; Evanston, IL, USA.
26. Moradi M, Shafiee Sardasht M, Ebrahimpour M. Bankruptcy prediction by support vector machines and multiple discriminate analysis models. Journal of Scurities Exchang 2012; 18(5): 113-36. [In Persian].
27. Sutton O. Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction [Online]. [cited 2012 Feb]; Available from: URL:http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf
28. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002; 46(1-3): 389-422.
29. Refaielzadeh P, Tang L, Liu H. Cross validation. Proceedings of AAAI Workshop on Evaluation Methods for Machine Learing II; 2007 Jul 22-23; Vancouver, Canada.
30. Drozdov I, Ouzounis CA, Shah AM, Tsoka S. Functional Genomics Assistant (FUGA): a toolbox for the analysis of complex biological networks. BMC Res Notes 2011; 4: 462.
31. Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol 2003; 4(10): R70.

Journal of Isfahan Medical School

Vol. 33, No. 359, 3rd Week, January 2016

Received: 02.10.2015

Accepted: 26.11.2015

Dimensionality Reduction on Topological Features of the Gene Network Constructed from Microarray Data for Prediction of Breast Cancer Recurrence

Alireza Mehridehnavi PhD[1], Hamed Zand[2], Mohammadreza Sehhati PhD[3]

Original Article

Abstract

Background: Extracted features from gene expression profiles of DNA microarrays are traditional tools in cancer classification. In this regard, using topological properties of genes through the gene network reconstruction can provide more reliable findings. The main goal of this article is the prediction of breast cancer recurrence via using topological features of the relevance network reconstructed from gene expression profiles.

Methods: We utilized seven gene expression microarray datasets, including 1271 samples from seven studies on breast cancer. In this study, the relevance gene network was reconstructed and FDA (Fisher Discriminant Analysis) method was applied for gene selection based on the characteristics of the network topology. To construct the gene network, we needed a profile of expressions for each gene and it could not be obtained from a single sample. Therefore, to classify a test sample, this sample was added to the training data and new gene networks were reconstructed according to two groups of high- and low-risk samples. The correlation coefficient between topological quantity vectors of the networks before and after adding test sample was calculated. The test sample was classified to the group that corresponded to higher correlation between new reconstructed network and the primary labeled network.

Findings: The classification accuracy was calculated using 5-fold cross-validation based on both correlation threshold and k-nearest neighbor (kNN) classifier and non-linear support vector machines (SVM) classifier that were applied on the topological properties of reconstructed gene networks. The results confirmed the advantage of applying topological features to the kNN and the non-linear SVM classifiers. The highest accuracy in prediction with the kNN classifier was obtained via degree centrality property that reached 98.5% in average among various numbers of genes.

Conclusion: Topological features of reconstructed gene networks from gene expression profiles provided more stable and accurate results in prediction of breast cancer recurrence.

Keywords: Breast cancer, Gene expression, Gene regulatory networks, Topology

Citation: Mehridehnavi A, Zand H, Sehhati M. **Dimensionality Reduction on Topological Features of the Gene Network Constructed from Microarray Data for Prediction of Breast Cancer Recurrence.** J Isfahan Med Sch 2016; 33(359): 1973-85

1- Associate Professor, Department of Biomedical Engineering (Bioelectronics), School of Advanced Medical Technology, Isfahan University of Medical Sciences, Isfahan, Iran

2- MSc Student, Department of Biomedical Engineering (Bioelectronics), School of Advanced Medical Technology AND Student Research Committee, Isfahan University of Medical Sciences, Isfahan, Iran

3- Assistant Professor, Department of Biomedical Engineering, School of Advanced Medical Technology, Isfahan University of Medical Sciences, Isfahan, Iran

Corresponding Author: Mohammadreza Sehhati PhD, Email: mr.sehhati@gmail.com