

به کارگیری روش‌های خوشبندی در ریزآرایه DNA

حمید علوی مجد^{*}، محسن واحدی^{**}، یدالله محرابی^{***}، بهار نقوی^{****}

* دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی

** گروه آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی

*** دانشکده بهداشت، دانشگاه علوم پزشکی شهید بهشتی

**** دانشکده پرستاری و مامایی، دانشگاه علوم پزشکی شهید بهشتی

چکیده

سابقه و هدف: به کارگیری فناوری ریزآرایه DNA که امکان بررسی بیان هزاران ژن را به طور همزمان در حداقل زمان ممکن می‌سازد، در سال‌های اخیر موجب تولید حجم انبوحی از داده‌های بیان ژنی شده است. تحلیل آماری این داده‌ها شامل مواردی چون نرم‌افزاری، خوشبندی، طبقه‌بندی است. هدف این مقاله بررسی نحوه به کارگیری روش‌های خوشبندی در داده‌های ریزآرایه DNA است.

روش بررسی: در این تحقیق داده‌های سرطان پستان وانتور و همکاران (۲۰۰۲) مربوط به جهش‌های ژنتیکی BRCA1 و BRCA2 تحلیل شده است. مجموعه داده‌ها شامل ۱۸ بیمار با جهش BRCA1 و ۲ بیمار با جهش BRCA2 است. داده‌های بیان ژنی سرطان پستان با استفاده از روش‌های آماری سلسله مراتبی و غیر سلسله مراتبی خوشبندی گردید. در هر دو روش خوشبندی، داده‌ها به دو خوشبندی تقسیم شدند. روش‌های مختلف خوشبندی با توجه به گروه‌بندی واقعی (BRCA2/BRCA1) مقایسه شدند. نرم‌افزار R برای تحلیل داده‌ها استفاده شد.

یافته‌ها و بیان: ویژگی روش خوشبندی سلسله مراتبی در تشخیص ژن BRCA1 ۹۴ درصد و حساسیت آن ۱۰۰ درصد بدست آمد. ویژگی روش خوشبندی غیر سلسله مراتبی در تشخیص ژن BRCA1 ۸۹ درصد و حساسیت آن ۱۰۰ درصد بدست آمد که نشان‌دهنده عملکرد مناسب دو روش خوشبندی است. روش خوشبندی سلسله مراتبی بر اساس ادغام بر حسب میانگین مناسب ترین روش در بین همه روش‌های بررسی شده است. نمونه شماره ۹۵ طبق نتایج همگی روش‌های خوشبندی در گروه BRCA2 قرار گرفت در صورتی که بر اساس یافته‌های بالینی در گروه BRCA1 قرار دارد.

نتیجه‌گیری: با توجه به انطباق قابل توجه نتایج خوشبندی با گروه‌بندی واقعی داده‌ها، می‌توان از این روش‌های آماری در مواردی که اطلاع دقیقی از گروه‌بندی واقعی داده‌ها در دست نیست، استفاده کرد. به علاوه نتایج خوشبندی ممکن است زیر گروه‌هایی از نمونه‌ها را به نحوی تمایز کند که برای انطباق آن با یافته‌های بالینی، پژوهش‌های آزمایشگاهی یا بالینی جدیدی لازم باشد.

وازگان کلیدی: ریزآرایه DNA، خوشبندی، بیان ژن، سرطان پستان

مقدمه

دومین علت مرگ و میر بهشمار می‌رود. در کشورهای در حال توسعه نیز بیماری سرطان در ردیف مسايل مهم بهداشتی درمانی محسوب و روند آن رو به افزایش است (۱).

سرطان پستان شایع‌ترین سرطان زنان در سراسر دنیا بوده و علت عده مرگ‌های ناشی از سرطان در زنان است. محققین در مطالعات مختلف نقش عوامل متعدد خطرسازی را برای سرطان پستان شناسایی کرده‌اند. شناخت و کشف علتها به جلوگیری از بیماری در افراد سالم و نسل‌های بعد کمک

در حال حاضر سرطان یکی از مسايل مهم و اصلی بهداشت و درمان در ایران و تمام دنیا می‌باشد. در آمریکا و تعدادی از کشورهای اروپایی، این بیماری بعد از بیماری‌های قلبی-عروقی

آدرس نویسنده مسئول: تهران، دانشگاه علوم پزشکی شهید بهشتی، گروه آمار زیستی دانشکده پیراپزشکی، دکتر حمید علوی مجد (email: alavimajd@gmail.com)

تاریخ دریافت مقاله: ۱۳۸۵/۷/۳۰

تاریخ پذیرش مقاله: ۱۳۸۶/۱/۳۱

به کارگیری روش‌های خوشبندی در ریزآرایه DNA

هیبرید می‌شوند. دو نوع آرایه بیشترین کاربرد را دارند:
۱- آرایه‌های بر پایه DNA مکمل (DNA complementary spotted)
۲- آرایه الیگونوکلئوتید (Oligonucleotide array) که به اختصار الیگو گفته می‌شود (۸).

در روش آرایه cDNA هر ژن با یک رشته طولانی (بین ۲۰۰ تا ۵۰۰ پایه) نشان داده می‌شود. cDNA از دو نمونه متفاوت بدست می‌آید، یکی از نمونه‌ی مورد آزمون و دیگری نمونه‌ی مرجع که بر روی یک آرایه هیبرید (مخلوط) می‌شوند. نمونه آزمون با رنگ فلورسنت قرمز و نمونه مرجع با رنگ سبز علامت گذاری می‌شوند. سپس با هدف تحریک رنگ‌های فلورسنت در دو طول موج مختلف آرایه بهوسیله اشعه لیزر اسکن می‌شود. از هر کدام این رنگ‌ها یک تصویر بوجود می‌آید که این تصاویر در کامپیوتر بر روی هم قرار داده می‌شود که حاصل این کار تراشه‌ای خواهد بود که حاوی هزاران لکه رنگی با رنگ‌های بسیار متنوع است که از ترکیب دو رنگ قرمز و سبز حاصل شده است. اندازه بیان ژنی می‌تواند لگاریتم نسبت شدت رنگ قرمز به سبز باشد (۹-۱۱).

در روش آرایه الیگونوکلئوتید هر ژن به ۱۶ الی ۲۰ حالت نشان داده می‌شود که هر کدام خود توالی کوتاهی از نوکلئوتیدها هستند و یک جفت کامل (Perfect Match) یا PM از یک قطعه ژن می‌باشد، در مقابل این ۲۰ الیگونوکلئوتید، ۲۰ الیگونوکلئوتید دیگر وجود دارد که به جز در باز مرکزی توالی آنها با هم برابر است، که به این نوکلئوتیدها غیرجفت صورت متوسط شدت اختلافات در این ۱۶ تا ۲۰ حالت می‌باشد (۸،۱۲).

برای بررسی داده‌های ریزآرایه DNA می‌توان از نرمافزارهای آماری نظری SAS ، S-plus ، STATA و R استفاده کرد. R به علت توانایی بالا در کار کردن با داده‌های حجمی رواج بیشتری دارد (۱۳).

مواد و روشها

از آنجایی که ریزآرایه DNA و روش‌های تحلیل داده‌های به دست آمده از آن جدید می‌باشد، تاکنون در بسیاری از کشورها از جمله ایران کار زیادی برای تولید این گونه داده‌ها صورت نگرفته است، بنابراین اغلب محققان ناچارند که از داده‌های بانک‌های اطلاعاتی اینترنتی استفاده کنند.

خواهد نمود. از جمله عواملی که در بروز سرطان پستان مؤثر است، عوامل ژنتیکی و عوامل محیطی است (۲). همچنین طبقه‌بندی دقیق و قابل اعتمادی از تومورها باید تعیین گردد تا مراحل درمانی مناسب با نوع خاص سرطان پستان در نظر گرفته شود. این طبقه‌بندی‌ها اغلب بر اساس شواهد بالینی و مشخصات مورفولوژی سلولی انجام می‌گیرد. ولی با ظهور فناوری ریزآرایه (Microarray) و استفاده از آن در این بخش از سرطان‌ها انتظار می‌رود که با تحلیل آماری تغییرات بیان هزاران ژن به طور همزمان بتوان فرایندهای ایجاد سرطان را شناسایی و در زمینه درمان این بیماری گام‌های مهمی برداشت. جهش‌های ژنتیکی BRCA1 و BRCA2 از عوامل اصلی به وجود آمدن سرطان پستان است. این دو نوع جهش بسیار شبیه هستند و برای شناسایی آنها از بدیگر نیاز به آزمایش‌های دقیق است. با توجه به شناسایی کامل ژنوم انسانی و همچنین ژن‌هایی که در این دو جهش نقش دارند می‌توان انتظار داشت تشخیص بین این دو جهش با استفاده از روش‌هایی بر مبنای بیان ژنی امکان‌پذیر باشد. با توجه به اینکه به کارگیری فناوری ریزآرایه DNA در سال‌های اخیر موجب تولید حجم انبوهی از داده‌های بیان ژنی شده است، یکی از روش‌های آماری که در تحلیل این داده‌ها به صورت فرایندهای مورد استفاده قرار می‌گیرد خوشبندی است. خوشبندی می‌تواند در یافتن گروههای واقعی و نهفته در داده‌ها موثر باشد، باعث کاهش داده‌ها شده و منجر به کشف خوشبندی جدید و غیرقابل انتظار شود. نرمافزاری برای تحلیل و ترسیم بر پایه الگوریتم‌های خوشبندی سلسله مراتبی تهیه گردیده است (۳). الیگواری (Oligoarray) به عنوان تشخیص‌دهنده ژنتیکی برای طبقه‌بندی بین دو نوع سرطان خون حاد (لفوئیدی و میلوئیدی) استفاده گردیده است (۴). مطالعه تومورهای بدخیم پوستی (جلدی) با استفاده از داده‌های بیان ژنی آرایه cDNA آنالیز خوشبندی صورت گرفت و زیر مجموعه‌ای ناشناخته از این بیماری تشخیص داده شد (۵). داده‌های بیان ژنی بافت‌های نمونه سرطان تخمدان خوشبندی گردید (۶). همچنین داده‌های بیان ژنی هورمون تیروئید بررسی شد و مجموعه‌ای از ژن‌ها بر اساس کارکردشان خوشبندی گردید (۷).

ریزآرایه ابزاری برای اندازه‌گیری و کسب اطلاعات از بیان ژن‌هاست. هر توالی ژنی شناخته شده مورد نظر به عنوان یک پروب (Prob) روی یک آرایه (Array) شیشه‌ای یا نایلونی چاپ می‌شود. mRNA از بافت یا نمونه خون با رنگ‌های فلورسنت علامت گذاری می‌شود و پروب‌ها بر روی یک آرایه

$$X = \begin{bmatrix} x_{11} & x_{12} & \Lambda & x_{1p} \\ x_{21} & x_{22} & \Lambda & x_{2p} \\ M & M & O & M \\ x_{n1} & x_{n2} & \Lambda & x_{np} \end{bmatrix} \xrightarrow{\text{transformed}} D_{n \times n}$$

عنصر x_{ik} در ماتریس X نشان دهنده بیان ژن i در فرد k است.

ضریب همبستگی پیرسون (Pearson's correlation coefficient) به عنوان معیار شباهت استفاده گردید که به صورت

$$\rho_{ij} = \frac{\sum_{k=1}^p (x_{ik} - \bar{x}_{i.})(x_{jk} - \bar{x}_{j.})}{\sqrt{\sum_{k=1}^p (x_{ik} - \bar{x}_{i.})^2 \sum_{k=1}^p (x_{jk} - \bar{x}_{j.})^2}}$$

در نظر گرفته می‌شود. با توجه به ضریب همبستگی پیرسون معیار اختلاف $d_{ij} = 1 - \rho_{ij}$ در نظر گرفته می‌شود که اختلاف بین نمونه i و j است و همچنین d_{ij} عنصر سطر i و ستون j ماتریس D می‌باشد.

این تکنیک خوشبندی خود دو نوع است: تجمعی (Agglomerative) و تقسیمی (Devision). نتایج هر دوی این روش‌ها قطعی و برگشت‌ناپذیر است. بدین مفهوم که در روش تجمعی اگر دو آزمودنی در یک خوش قرار گرفتند، دیگر قابل تفکیک نخواهند بود و با اگر در روش تقسیم در دو خوشه مجزا قرار گرفتند، دیگر امکان اینکه در یک خوش با هم واقع شوند وجود ندارد.

نتایج حاصل از روش خوشبندی سلسله مراتبی در نمودارهای دندوگرام (Dendrogram) یا نمودار درختی نمایش داده می‌شود. در دندوگرام‌ها محور عمودی فاصله بین خوشه‌ها را اندازه‌گیری می‌کند و ارتفاع هر یک از شاخه‌ها بیانگر آن است که دو خوشه مورد نظر در چه نقطه‌ای با هم ادغام شده‌اند (۱۶). برای مقایسه نمودارهای درختی از معیار فاصله کوفنتیک (Cophenetic distances) به منظور انعکاس فواصل واقعی در نمودار درختی استفاده کردیم. هر نموداری که معیار فاصله کوفنتیک آن بالاتر باشد خوشبندی داده‌ها را بهتر نشان داده است (۱۷).

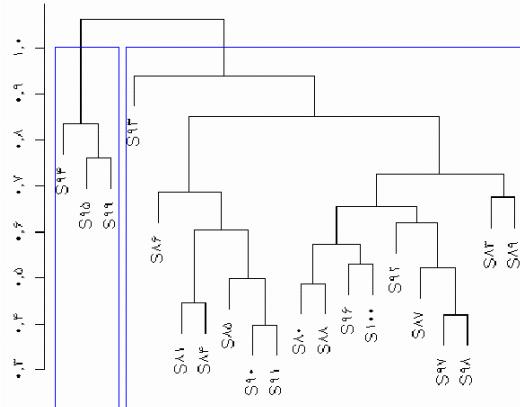
روش‌های خوشبندی غیر سلسله مراتبی (Non-hierarchical) برای دسته‌بندی کردن اقلام بجای متغیرها به مجموعه‌ای از خوشه طراحی شده است. یکی از روش‌های غیر سلسله مراتبی

در این تحقیق از داده‌های سرطان پستان که توسط وانتور و همکاران در سال ۲۰۰۲ انتشار یافت، استفاده شد (۱۴). مقاله وانتور شامل ۴ مجموعه داده است، یکی از این مجموعه داده‌ها ArrayData_BRCA1 است که در این مقاله از آن استفاده شده است. این مجموعه شامل ۱۸ بیمار دارای جهش BRCA1 و ۲ نفر بیمار دارای جهش BRCA2 است که برای هر کدام از آنها مقدار بیان ژنی ۲۴۴۲۵ ژن جمع‌آوری گردیده است. این داده‌ها در اینترنت در اختیار عموم قرار دارد (۱۵).

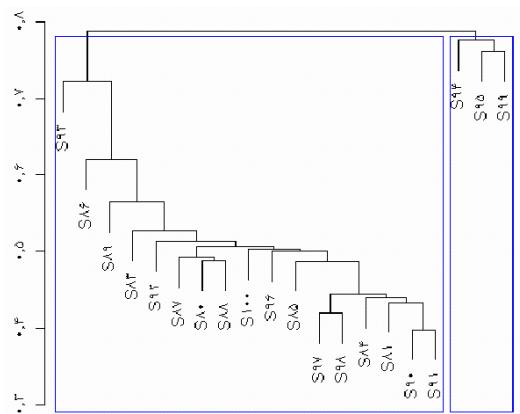
سوال مشترک در مطالعات بیان ژنی که بر اساس آنها از روش‌های تحلیل آماری متفاوتی استفاده می‌شود، اغلب در یکی از این موارد قرار دارد: ۱- کشف طبقات (Class discovery)، ۲- تشخیص ژنی (Gene identification)، ۳- پیش‌بینی طبقات (Class prediction). کشف طبقات شناسایی زیرگروه‌های نامعلوم است، در حالی که روش‌های پیش‌بینی طبقات برای طبقه‌بندی نمونه مستقل جدید بر اساس یک طرح پیش‌بینی مورد استفاده قرار می‌گیرند. تشخیص ژنی مربوط به شناسایی ژن‌هایی است که بیان آنها دچار تغییر شده است (۸). در این تحقیق ما به سؤوال اول با استفاده از روش خوشبندی جواب می‌دهیم.

خوشبندی (Clustering) در واقع تقسیم‌بندی یک جمعیت ناهمگون (Heterogeneous population) به تعدادی از زیرمجموعه‌های همگون (Homogeneous) می‌باشد که به آنها خوش اطلاق می‌شود. هدف از خوشبندی، یافتن گروه‌هایی است که با یکدیگر بسیار متفاوتند ولی اعضای این گروه‌ها بسیار شبیه به هم هستند. به عبارت دیگر در خوشبندی مجموعه‌ای از عناصر یا افراد را بر مبنای مشابهت آنها به چند زیرمجموعه کوچک‌تر تقسیم می‌کنیم، به طوری که بین عناصر در یک خوش بیشترین شباهت و بین عناصر واقع در دو خوشه مجزا بیشترین تمایز وجود داشته باشد (۱۶). در این تحقیق از دو روش سلسله مراتبی و غیر سلسله مراتبی برای خوشبندی استفاده گردید.

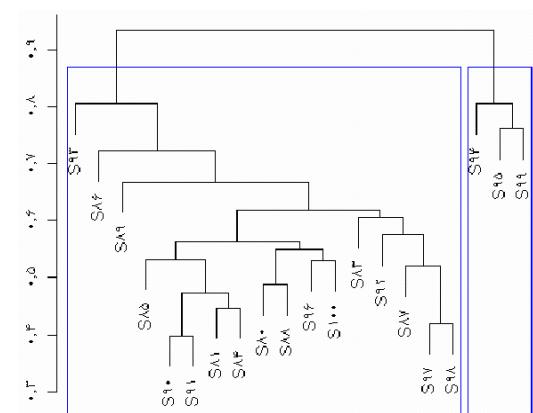
روش خوشبندی سلسله مراتبی (Hierarchical) با یک سری ادغام‌ها یا تقسیمات متوالی از آزمودنی‌ها همراه است. اساس تعیین خوشه در این روش محاسبه اندازه شباهت‌ها (Similarity) و یا اختلاف‌ها (Distance) بین هر زوج از عناصر مورد مطالعه است. به این ترتیب بر اساس ماتریس شباهت یا اختلاف (D) که بر اساس ماتریس داده‌ها (X) به دست آمده خوشبندی انجام گردید.



نمودار ۲- نمودار درختی با روش ادغام بر اساس دورترین همسایه‌ها(کامل)



نمودار ۳- نمودار درختی با روش ادغام بر اساس نزدیکترین همسایه‌ها(منفرد)



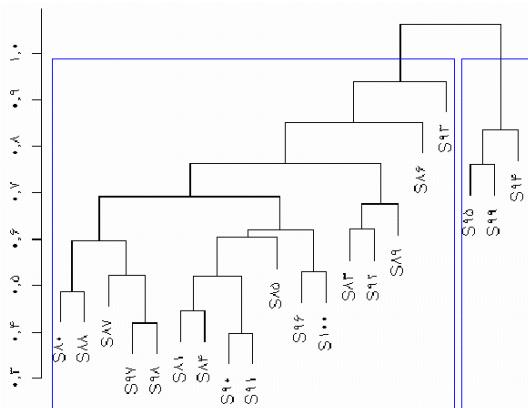
نمودار ۴- نمودار درختی با روش سلسله مراتبی تقسیمی

که در این تحقیق ما از آن استفاده کردیم، روش k میانگین (k-means) است. معمولاً در عمل برای پیدا کردن مقدار مناسب k ابتدا یک خوشبندی سلسله مراتبی بر روی داده‌ها انجام می‌گیرد و مقدار اولیه k تعیین می‌گردد (۸). بنا به روش خوشبندی سلسله مراتبی که در داده‌های این تحقیق انجام گردید، مقدار اولیه $k=2$ در نظر گرفته شد. با توجه به معیار هارتیگان اگر نتیجه فرمول زیر بزرگ‌تر از 10 باشد به اضافه کردن خوشبندی ادامه می‌دهیم تا به ازای اولین مقدار k ای که به ازای آن این عبارت از 10 کوچک‌تر شد آنرا به عنوان مقدار مناسب k انتخاب می‌شود (۱۸). مجموع مربعات درون خوشبندی وقتی k خوشه داریم و n تعداد نمونه‌هاست).

$$\frac{SS_k}{SS_{k+1}-1} \times (n-k-1)$$

یافته‌ها

با انجام پیش‌پردازش (Preprocessing) اولیه ۲۴۱۸۸ ژن باقی ماندند. قابل ذکر است که قبل از تحلیل این نوع داده‌ها لازم است مواردی مانند چولگی، اریبی و مقادیر دور افتاده (Outliers) بررسی و در صورت وجود تصحیح شود، که به آن پیش‌پردازش گفته می‌شود. در اولین مرحله ماتریس فاصله 20×20 که هر عضو آن مقدار یک منهای ضربی همبستگی پیرسون بین بیماران بر مبنای ۲۴۱۸۸ ژن باقیمانده است، تشکیل گردید. نتایج خوشبندی سلسله مراتبی در نمودارهای ۱ تا ۴ نشان داده شده است.



نمودار ۱- نمودار درختی با روش ادغام بر اساس مراکز خوشبندی (میانگین)

نمونه ۲ BRCA2 را درست تشخیص داده است. با توجه به جدول مذکور ویژگی روش خوشبندی سلسله مراتبی در تشخیص افراد BRCA1، ۹۴ درصد و حساسیت آن ۱۰۰ درصد بهدست آمد که نشان دهنده عملکرد مناسب روش خوشبندی سلسله مراتبی است.

در خوشبندی غیر سلسله مراتبی با استفاده از روش k میانگین، ابتدا با $k=2$ شروع کردیم. نتایج روش k میانگین برای $k=2$ و $k=3$ در جدول ۲ آمده است. از جدول ۳ مجموع مربعات درون خوشبندی است، برای $k=2$ و $n=20$ معیار هارتیگان برابر $8/7511$ بهدست آمد که کوچکتر از ۱۰ بوده، بنابراین ۲ خوشبندی مناسب می‌باشد. نتایج این روش در جدول ۴ آمده است. با توجه به این جدول ویژگی روش خوشبندی غیرسلسله مراتبی در تشخیص افراد BRCA1، ۸۹ درصد و حساسیت آن ۱۰۰ درصد بدست آمد.

جدول ۳- مجموع مربعات درون خوشبندی

$k=3$	$k=2$	
۰/۰۰۰	۲۷۶۳/۶۰۱	خوشبندی ۱
۸۹۷۸/۱۲۹	۲۴۲۹/۶۲۵	خوشبندی ۲
۱۱۰۵/۴۷۷	_____	خوشبندی ۳

جدول ۴- انتساب بیماران با روش خوشبندی غیرسلسله مراتبی (k میانگین)

جهش ژنتیکی	خوشبندی ۱	خوشبندی ۲	خوشبندی ۳	جمع	$k=3$	$k=2$	نمونه
BRCA1	۱۶	۲	۲	۲۰	۱۸	۱	S80
BRCA2	۰	۲	۲	۴	۲	۲	S81
جمع	۱۶	۴	۴	۲۰	۲۰	۳	S83

بحث

در این تحقیق داده‌های بیان ژنی سرطان پستان با استفاده از روش‌های آماری (خوشبندی سلسله مراتبی و غیر سلسله مراتبی) خوشبندی گردید. با مقایسه بین معیار فاصله کوفنتیک مشخص شد که روش خوشبندی سلسله مراتبی بر اساس ادغام بر حسب میانگین روش مناسبی نسبت به دیگر روش‌های سلسله مراتبی است. هر اندازه که تعداد خوشبندی بیشتر شود روش خوشبندی سلسله مراتبی ادغام بر اساس دورترین همسایه‌ها زیرگروه‌های مجزایی از BRCA1 را ارایه

جدول ۱- انتساب بیماران با روش خوشبندی سلسله مراتبی

جهش ژنتیکی	خوشبندی ۱	خوشبندی ۲	جمع
BRCA1	۱۷	۱	۱۸
BRCA2	۰	۲	۲
جمع	۱۷	۳	۲۰

با توجه به معیار فاصله کوفنتیک که به ترتیب $0/92$ و $0/93$ برای حالت‌های ادغام بر حسب میانگین، ادغام کامل، ادغام منفرد و روش سلسله مراتبی تقسیمی است و با توجه به این نمودارها نمونه‌ها را می‌توان در دو خوشبندی کرد. خوشبندی اول شامل ۱۷ نمونه و خوشبندی دوم شامل ۳ نمونه است که نتیجه آن در جدول ۱ آمده است. روش خوشبندی سلسله مراتبی در تشخیص نمونه‌ها به دو

جدول ۲- تشخیص نمونه‌ها به دو و سه خوشبندی

جهش ژنتیکی	خوشبندی ۱	خوشبندی ۲	خوشبندی ۳	نمونه
BRCA1	۱	۲	۲	S80
BRCA2	۰	۱	۲	S81
جمع	۱۶	۴	۲	S83
BRCA1	۱۶	۲	۲	S84
BRCA2	۰	۱	۱	S85
جمع	۱۶	۱	۱	S86
BRCA1	۱۶	۱	۱	S87
BRCA2	۰	۱	۱	S88
جمع	۱۶	۱	۱	S89
BRCA1	۱۶	۱	۱	S90
BRCA2	۰	۱	۱	S91
جمع	۱۶	۱	۱	S92
BRCA1	۱۶	۱	۱	S93
BRCA2	۰	۲	۲	S95
جمع	۱۶	۱	۱	S96
BRCA1	۱۶	۱	۱	S97
BRCA2	۰	۱	۱	S98
جمع	۱۶	۱	۱	S100
BRCA1	۱۶	۲	۲	S94
BRCA2	۰	۲	۲	S99

خوشبندی اول از هیجده نمونه BRCA1 هفده نمونه و خوشبندی دوم هر دو

یافته‌های بالینی BRCA1 قرار گرفته است، احتمالاً شباهت ژنتیکی قابل توجه با گروه BRCA2 دارد. بنابراین می‌توان توصیه کرد، این گونه نمونه‌ها مورد آزمایشات بالینی بیشتری قرار گیرند تا از نحوه انتساب آنها به هر یک از دو گروه اطمینان بیشتری حاصل گردد. برازما و ویلو (۲۰۰۰) روش k میانگین را برای خوشبندی یک ماتریس بین ژنی 6121×80 به کار برند (۲۲). تقیزاده جهرمی و همکاران (۱۳۸۴) از مدل شطرنجی (Plaid models) برای خوشبندی داده‌های لوسمی میلئید حاد استفاده کردند (۲۳).

انطباق قابل توجه نتایج خوشبندی با گروه‌بندی واقعی داده‌ها، باعث گردیده از این روشنایی آماری در مواردی که اطلاع دقیقی از گروه‌بندی واقعی داده‌ها در دست نیست، استفاده شود و این روشنایی را می‌توان در کنار روشنایی بالینی به کار برد تا به اطمینان بیشتری به تشخیص بیماری دست یافت. در حقیقت این فناوری فرصتی را جهت افزایش دقیقت تجربیات بالینی فراهم کرده که می‌توان در سرطان‌هایی که به خوبی مطالعه شده‌اند به ارزیابی این تجربیات قبل از به کار بردن آن‌ها در بیماران پرداخت.

می‌دهد که برای بررسی‌های بیشتر شناسایی چنین زیرگروه‌هایی ضروری است. در بین روشنایی خوشبندی سلسله مراتبی روش خوشبندی سلسله مراتبی بر اساس ادغام نزدیکترین همسایه‌ها ضعیفترین نتایج را دارد. هدنفالک و همکاران (۲۰۰۱) گروه‌های مختلفی از ژن‌های بیان شده با جهش‌های BRCA1 و BRCA2 را گروه‌بندی کردند (۱۹). وی و همکاران (۲۰۰۴) با استفاده از خوشبندی سلسله مراتبی الگوی بیان ژنی بیماران در زیرگروه همگن را شناسایی کردند (۲۰). همچنین شوچی و همکاران (۲۰۰۶) به خوشبندی سلسله مراتبی ژن‌ها پرداختند و ژن‌های با بیان متفاوت را شناسایی کردند (۲۱).

با توجه به جدول‌های ۱ و ۴ روشنایی خوشبندی سلسله مراتبی در انتساب نمونه‌ها به ۲ خوشبندی عملکرد بهتری داشت، زیرا تنها یک انتساب اشتباه بود، در حالیکه روش k میانگین دارای ۲ انتساب اشتباه بود. از آنجا که در روش خوشبندی غیر سلسله مراتبی لازم نیست ماتریس فواصل (مشابهت‌ها) تعیین شود و داده‌های اصلی در طول اجرا ذخیره شوند، لذا روشنایی غیر سلسله مراتبی را می‌توان به مجموعه داده‌های بسیار بزرگ‌تری نسبت به روشهای سلسله مراتبی به کار برد. طبق نتایج خوشبندی نمونه شماره ۹۵ که بر اساس

REFERENCES

1. Haskell CM, Bereck J. Cancer Treatment. 5th ed. Philadelphia: WB Saunder; 2001.
2. Aghassi-Ippen M, Green MS. Familal risk factors for breast cancer among Arab women in Israel. Eur J Cancer Prev 2002; 11: 327–31.
3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 1998; 95: 14863–8.
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286: 531–7.
5. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000; 406: 536–40.
6. Welsh JB, Zarrinkar PP, Sapino LM, Kern SG, Behling CA, Monk BJ, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proceedings of the National Academy of Sciences 2001; 98: 1176–81.
7. Clement K, Viguerie N, Diehn M, Alizadeh A, Baribe P, Thalamas C, et al. In vivo regulation of human skeletal muscle gene expression by thyroid hormone. Genome Research 2002; 12: 281–91.
8. Satagopan JM, Panageas KS. Tutorial in biostatistics: a statistical perspective on gene expression data analysis. Stat Med 2003; 22: 481–99.
9. Eisen MB, Brown PO. DNA arrays for analysis of gene expression. Methods Enzymol 1999; 303: 179–205.
10. Amarasinghe D, Cabrera J. Exploration and analysis of DNA microarray and protein array data. New York: Wiley & Sons; 2004.
11. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Optic 1997; 2: 364–74.
12. Affymetrix Microarray Suite User Guide. Version 4.0. 2000; Appendix A2, A3.

13. Department of statistics and mathematics of the WU Wien. The R project for statistical computing. Available at: URL: <http://www.r-project.org>
14. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; 415: 530–6.
15. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Available at: URL: <http://www.rii.com/publications/2002/vantveer.html>
16. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. 4th edition. Upper Saddle River: Prentice-Hall; 1992.
17. Sneath PH, Sokal RR. The principles and practice of numerical classification. Numerical Taxonomy. San Francisco: W H Freeman & Co.; 1973: 278.
18. Hartigan JA. Clustering Algorithms. New York: Wiley & Sons; 1975.
19. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene expression profiles in hereditary breast cancer. *N Engl J Med* 2000; 34(48): 539-48.
20. Vey N, Mozziconacci MJ, Groulet-Martinec A, Debono S, Finetti P, Carbuccia N, et al. Identification of new classes among acute myelogenous leukemias with normal karyotype using gene expression profiling. *Oncogene* 2004; 23: 9381-91.
21. Guo SJ, Wu LY, Shen WL, Chen WD, Wei J, Gao PJ, et al. Gene profile for differentiation of vascular adventitial myofibroblasts. *Acta Physiologica Sinica* 2006; 8(4): 337-44.
22. Brazma A, Vilo J. Gene expression data analysis. *Federation of European Biochemical Societies Letters* 2000; 480: 17-24.
۲۳. تقیزاده جهرمی م. به کارگیری مدل شطرنجی در خوشه بندی دهنده های بیان ژنی لوسومی میلوقید حاد. پایان نامه کارشناسی ارشد، دانشکده پزشکی، دانشگاه تربیت مدرس، سال ۱۳۸۴.