پژوهش درپزشکی (مجله پژوهشی دانشکده پزشکی)
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهید بهشتی
دوره ۳۱، شماره ۱، بهار ۱۳۸۶، صفحات ۱۹ تا ۲۵

مقاله اصلی
(Original Article)

به کارگیری روش‌های خوشه‌بندی 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: *به‌کارگیری فناوری ریزآرایه DNA که امکان بررسی بیان هزاران ژن را به‌طور هم‌زمان در حداقل زمان ممکن می‌سازد، در سال‌های اخیر موجب تولید حجم انبوهی از داده‌های بیان ژنی شده است. تحلیل آماری این داده‌ها شامل مواردی چون نرمال‌سازی، خوشه‌بندی، طبقه‌بندی است. هدف این مقاله بررسی نحوه به‌کارگیری روش‌های آماری خوشه‌بندی در داده های ریز آرایه 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 خوشه‌بندی گردید. در هر دو روش خوشه‌بندی، داده‌ها به دو خوشه تقسیم شدند. روش‌های مختلف خوشه‌بندی با توجه به گروه‌بندی واقعی (BRCA1،BRCA2) مقایسه شدند. نرم‌افزار R برای تحلیل داده‌ها استفاده شد.*

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

نتیجه‌گیری: *با توجه به انطباق قابل توجه نتایج خوشه‌بندی 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 خوشه‌بندی ممکن است زیر گروه‌هایی از نمونه‌ها را به نحوی متمایز کند که برای انطباق آن با یافته‌های بالینی، پژوهش‌های آزمایشگاهی یا بالینی جدیدی لازم باشد.*

واژگان کلیدی: *ریزآرایه DNA، خوشه‌بندی، بیان ژن، سرطان پستان*

مقدمه

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

سرطان پستان شایع‌ترین سرطان زنان در سراسر دنیا بوده و علت عمده 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

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

خواهد نمود. از جمله عواملی که در بروز سرطان پستان موثر 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) DNA 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 به‌وجود آمدن سرطان پستان است. این دو نوع جهش بسیار شبیه هستند و برای شناسایی آنها از یکدیگر 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 به‌کارگیری 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 فزاینده‌ای مورد استفاده قرار می‌گیرد خوشه‌بندی است. خوشه‌بندی می‌تواند در یافتن گروه‌های واقعی و نهفته در داده‌ها موثر باشد، باعث کاهش داده‌ها شده و منجر به کشف خوشه‌های جدید و غیرقابل انتظار شود. نرم‌افزاری برای تحلیل و ترسیم بر پایه الگوریتم‌های خوشه‌بندی سلسله مراتبی تهیه گردیده است (۳). الیگواری (Oligoarray) به عنوان تشخیص‌دهنده ژنتیکی برای طبقه‌بندی بین دو نوع سرطان خون حاد (لنفوئیدی و میلوئیدی) استفاده گردیده است (۴). مطالعه تومورهای بدخیم پوستی (جلدی) با استفاده از داده‌های بیان ژنی آرایه cDNA آنالیز خوشه‌ای صورت گرفت و زیر مجموعه‌ای ناشناخته از این بیماری تشخیص داده شد (۵). داده‌های بیان ژنی بافت‌های نمونه سرطان تخمدان خوشه‌بندی گردید (۶). هم‌چنین داده‌های بیان ژنی هورمون تیروئید بررسی شد و مجموعه‌ای از ژن‌ها بر اساس کارکردشان خوشه‌بندی 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 علامت‌گذاری می‌شود و پروب‌ها بر روی یک آرایه 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 (۱۱-۹).

در روش آرایه الیگونوکلئوتید هر ژن به۱۶ 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 (MisMatch) یا MM می‌گویند. یک اندازه از بیان ژنی به صورت متوسط شدت اختلافات در این ۱۶ تا ۲۰ حالت می‌باشد (۸،۱۲).

برای بررسی داده‌های ریزآرایه DNA می‌توان از نرم‌افزارهای آماری نظیر SAS ، S-plus ، STATA و R استفاده کرد. R به علت توانایی بالا در کار کردن با داده‌های حجیم 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.

در این تحقیق از داده‌های سرطان پستان که توسط وانت‌ور و همکاران در سال ۲۰۰۲ انتشار یافته، 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 گردیده‌است. 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 حالی‌که 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 سئوال اول با استفاده از روش خوشه‌بندی جواب می‌دهیم.

خوشه‌بندی (Clustering) در واقع تقسیم‌بندی یک جمعیت ناهمگون (Heterogeneous population) به تعدادی از زیرمجموعه‌های همگون(Homogeneous) می‌باشد که به آنها خوشه اطلاق می‌شود. هدف از خوشه‌بندی، یافتن گروه‌هایی است که با یکدیگر بسیار متفاوتند ولی اعضای این گروه‌ها بسیار شبیه به هم هستند. به عبارت دیگر در خوشه‌بندی مجموعه‌ای از عناصر یا افراد را بر مبنای مشابهت آن‌ها به چند زیر مجموعه کوچک‌تر تقسیم می‌کنیم، به‌طوری‌که بین عناصر در یک خوشه 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 خوشه‌بندی استفاده گردید.

روش خوشه‌بندی 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) به‌دست آمده خوشه‌بندی انجام گردید.

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

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

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

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

این تکنیک خوشه‌بندی 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 خوشه قرار گرفتند، دیگر قابل تفکیک نخواهند بود و یا اگر در روش تقسیم در دو خوشه مجزا قرار گرفتند، دیگر امکان اینکه در یک خوشه با هم واقع شوند وجود ندارد.

نتایج حاصل از روش خوشه‌بندی 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 (۸،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 خوشه‌بندی داده‌ها را بهتر نشان داده است (۱۷).

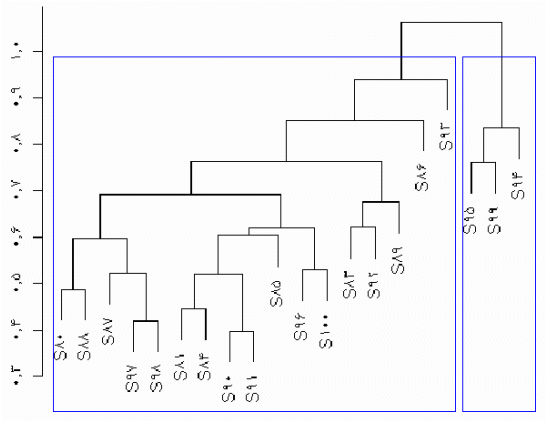
روش‌های خوشه‌بندی غیر سلسله مراتبی (Non-hierarchical) برای دسته‌بندی کردن اقلام بجای متغیرها به مجموعه‌ای از k خوشه طراحی شده است. یکی از روش‌های غیر سلسله مراتبی

که در این تحقیق ما از آن استفاده کرده‌ایم، روش k میانگین (k-means) است. معمولاً در عمل برای پیدا کردن مقدار مناسب k ابتدا یک خوشه‌بندی سلسله مراتبی بر روی داده‌ها انجام می‌گیرد و مقدار اولیه k تعیین می‌گردد (۸). بنا به روش خوشه‌بندی سلسله مراتبی که در داده‌های این تحقیق انجام گردید، مقدار اولیه k، ۲ در نظر گرفته شد. با توجه به معیار هارتیگان اگر نتیجه فرمول زیر بزرگتر از ۱۰ باشد به اضافه کردن خوشه‌ها ادامه می‌دهیم تا به ازای اولین مقدار k ای که به ازای آن این عبارت از ۱۰ کوچک‌تر شد آنرا به عنوان مقدار مناسب k انتخاب می‌شود (۱۸). (SS_k مجموع مربعات درون خوشه‌ای وقتی k خوشه داریم و n تعداد نمونه‌هاست.)

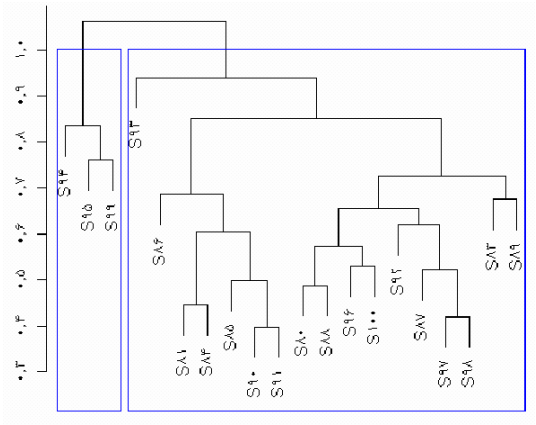
$$\frac{SS_k}{SS_{k+1}-1}\times(n-k-1)$$

یافته‌ها

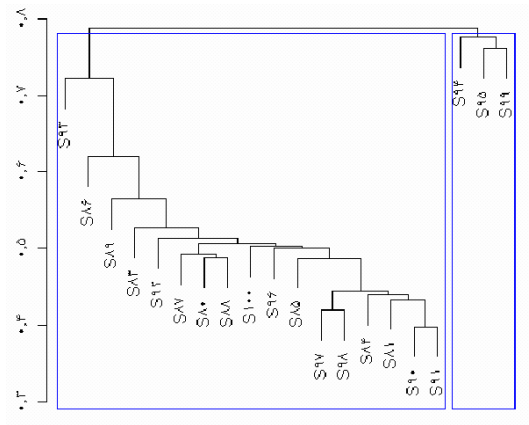
با انجام پیش‌پردازش (Preprocessing) اولیه ۲۴۱۸۸ ژن باقی ماندند. قابل ذکر است که قبل از تحلیل این نوع داده ها لازم است مواردی مانند چولگی، اریبی و مقادیر دور افتاده (Outliers) بررسی و در صورت وجود تصحیح شود، که به آن پیش پردازش گفته می‌شود. در اولین مرحله ماتریس فاصله ۲۰×۲۰ که هر عضو آن مقدار یک منهای ضریب همبستگی پیرسون بین بیماران بر مبنای ۲۴۱۸۸ ژن باقیمانده است، تشکیل گردید. نتایج خوشه‌بندی سلسله مراتبی در نمودارهای ۱ تا ۴ نشان داده شده است.



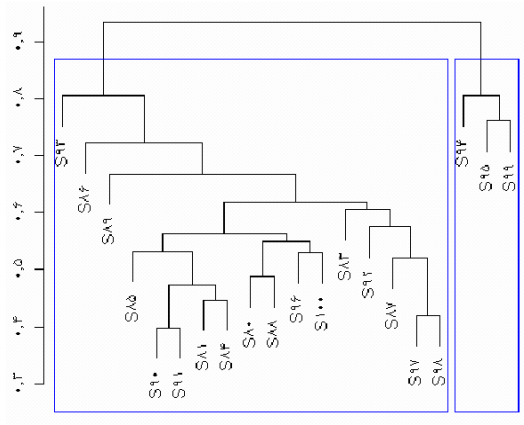
نمودار ۱- نمودار درختی با روش ادغام بر اساس مراکز خوشه‌ای(میانگین



نمودار۲- 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

جدول۱- انتساب بیماران با روش خوشه‌بندی سلسله مراتبی

جهش ژنتیکی	خوشه۱	خوشه۲	جمع
BRCA1	۱۷	۱	۱۸
BRCA2	۰	۲	۲
جمع	۱۷	۳	۲۰

با توجه به معیار فاصله کوفنتیک که به ترتیب ۰/۹۴، ۰/۸۵، ۰/۹۲ و ۰/۹۳ برای حالت‌های ادغام برحسب میانگین، ادغام کامل، ادغام منفرد و روش سلسله مراتبی تقسیمی است و با توجه به این نمودارها نمونه‌ها را می‌توان در دو خوشه دسته‌بندی کرد. خوشه اول شامل ۱۷ نمونه و خوشه دوم شامل ۳ نمونه است که نتیجه آن در جدول ۱ آمده است. روش خوشه‌بندی سلسله مراتبی در تخصیص نمونه‌ها به دو

جدول ۲- تخصیص نمونه‌ها به دو و سه خوشه

نمونه	k=2	k=3
S80	۱	۲
S81	۱	۲
S83	۲	۲
S84	۱	۲
S85	۱	۲
S86	۱	۲
S87	۱	۲
S88	۱	۲
S89	۱	۲
S90	۱	۲
S91	۱	۲
S92	۱	۲
S93	۱	۱
S95	۲	۲
S96	۱	۲
S97	۱	۲
S98	۱	۲
S100	۱	۲
S94	۲	۳
S99	۲	۳

خوشه BRCA1 و BRCA2 توانایی بالایی دارد، زیرا خوشه اول از هیجده نمونه BRCA1 هفده نمونه و خوشه دوم هر دو نمونه BRCA2 را درست تشخیص داده است. با توجه به جدول مذکور ویژگی روش خوشه‌بندی سلسله مراتبی در تشخیص افراد BRCA1، ۹۴ درصد و حساسیت آن ۱۰۰ درصد به‌دست آمد که نشان‌دهنده عملکرد مناسب روش خوشه‌بندی سلسله مراتبی است.

در خوشه‌بندی غیر سلسله مراتبی با استفاده از روش k میانگین، ابتدا با k=۲ شروع کردیم. نتایج روش k میانگین برای k=۲ و k=۳ در جدول ۲ آمده است. از جدول ۳ مجموع مربعات درون خوشه‌ای است، برای k=۲ و n=۲۰ معیار هارتیگان برابر ۸/۷۵۱۱ به‌دست آمد که کوچک‌تر از ۱۰ بوده، بنابراین ۲ خوشه مناسب می‌باشد. نتایج این روش در جدول ۴ آمده است. با توجه به این جدول ویژگی روش خوشه‌بندی غیرسلسله مراتبی در تشخیص افراد BRCA1، ۸۹ درصد و حساسیت آن ۱۰۰ درصد بدست آمد.

جدول۳- مجموع مربعات درون خوشه‌ای

	k=2	k=3
خوشه ۱	۲۷۶۳/۶۰۱	۰/۰۰۰
خوشه۲	۲۴۲۹/۶۲۵	۸۹۷۸/۱۲۹
خوشه۳	_____	۱۱۰۵/۴۷۷

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

جهش ژنتیکی	خوشه۱	خوشه۲	جمع
BRCA1	۱۶	۲	۱۸
BRCA2	۰	۲	۲
جمع	۱۶	۴	۲۰

بحث

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

می‌دهد که برای بررسی‌های بیشتر شناسایی چنین زیرگروه‌هایی ضروری است. در بین روش‌های خوشه‌بندی سلسله مراتبی روش خوشه‌بندی سلسله مراتبی بر اساس ادغام نزدیکترین همسایه‌ها ضعیف‌ترین نتایج را دارد. هدنفالک و همکاران (۲۰۰۱) گروه‌های مختلفی از ژن‌های بیان شده با جهش های BRCA1 وBRCA2 را گروه‌بندی کردند (۱۹). وی و همکاران (۲۰۰۴) با استفاده از خوشه بندی سلسله مراتبی الگوی بیان ژنی بیماران در زیر گروه همگن را شناسایی کردند (۲۰). همچنین شوجی و همکاران (۲۰۰۶) به خوشه‌بندی ساسله مراتبی ژن‌ها پرداختند و ژن های با بیان متفاوت را شناسایی کردند (۲۱).

با توجه به جدول‌های ۱ و ۴ روش‌های خوشه‌بندی سلسله مراتبی در انتساب نمونه‌ها به ۲ خوشه عملکرد بهتری داشت، زیرا تنها یک انتساب اشتباه بود، در حالیکه روش k میانگین دارای ۲ انتساب اشتباه بود. از آنجا که در روش خوشه‌بندی غیر سلسله مراتبی لازم نیست ماتریس فواصل (مشابهت‌ها) تعیین شود و داده‌های اصلی در طول اجرا ذخیره شوند، لذا روش‌های غیر سلسله مراتبی را می‌توان به مجموعه داده‌های بسیار بزرگ‌تری نسبت به روشهای سلسله مراتبی به کار برد. طبق نتایج خوشه‌بندی نمونه شماره ۹۵ که بر اساس یافته‌های بالینیBRCA1 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 خوشه بندی یک ماتریس بین ژنی ۸۰×۶۱۲۱ به کار بردند (۲۲). تقی‌زاده جهرمی و همکاران (۱۳۸۴) از مدل شطرنجی (Plaid models) برای خوشه‌بندی داده‌های لوسمی میلوئید حاد استفاده کردند (۲۳).

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

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, Sapinoso 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, Barbe 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. Amaratunga 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 Physioligica 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 دهنده‌های بیان ژنی لوسمی میلوئید حاد. پایان نامه کارشناسی ارشد، دانشکده پزشکی، دانشگاه تربیت مدرس، سال ۱۳۸۴.