مجله علمی دانشگاه علوم پزشکی سمنان- جلد ۹، شماره ۲ (پیاپی ۲۶) ، زمستان ۱۳۸۶

# خوشه‌بندی داده‌های بیان ژنی و کاربرد آن در تحلیل افتراق انواع سرطان خون

محسن واحدی[۱*](M.SC) ، حمید علوی مجد[۲] (Ph.D)، یدا... محرابی[۳] (Ph.D)، بهار نقوی[۴](M.SC)

۱ – *دانشگاه علوم پزشکی شهید بهشتی، مرکز تحقیقات بیماری‌های گوارش و کبد*

۲ – *دانشگاه علوم پزشکی شهید بهشتی، دانشکده پیرایزشکی*

۳ – *دانشگاه علوم پزشکی شهید بهشتی، دانشکده بهداشت*

۴ – *دانشگاه علوم پزشکی شهید بهشتی، دانشکده پرستاری*

## چکیده

**سابقه و هدف: یکی از شاخه‌های مهم بیوانفورماتیک فناوری ریزآرایه DNA است که امکان بررسی بیان هزاران ژن را به طور هم‌زمان در حداقل زمان ممکن می‌سازد که در سال‌های اخیر موجب تولید حجم انبوهی از داده‌های بیان ژنی شده است. تحلیل آماری این داده‌ها شامل نرمال سازی، خوشه بندی، طبقه بندی و ... از جمله روش‌های مورد استفاده در تحلیل این نوع داده‌ها است.**

**مواد و روش‌ها: در این مقاله داده‌های بیان ژنی سرطان خون گلوب و همکاران (۱۹۹۹) که بر اساس روش آرایه الیگونوکلئوتید تولید شده و از طریق اینترنت در اختیار عموم قرار دارد، با استفاده از روش‌های آماری مقیاس‌بندی چند بعدی، خوشه‌بندی سلسله مراتبی و غیر سلسله مراتبی مورد تجزیه و تحلیل قرار گرفته است. مجموعه داده‌ها شامل ۲۰ بیمار مبتلا به سرطان خون لنفوئیدی حاد (ALL) و ۱۴ بیمار مبتلا به سرطان خون میلوئیدی حاد (AML) است. در هر دو روش خوشه‌بندی، داده‌ها به دو خوشه تقسیم شدند. روش‌های مختلف خوشه‌بندی با توجه به گروه‌بندی واقعی نمونه‌ها (ALL،AML) مورد مقایسه قرار گرفتند. نرم افزار R برای تحلیل داده‌ها استفاده شد.**

**یافته‌ها: ویژگی روش خوشه‌بندی سلسله مراتبی تقسیمی در تشخیص افراد ALL، ۷۵ درصد و حساسیت آن ۹۲ درصد بدست آمد، ویژگی روش افراز کردن اطراف میدوئید در تشخیص افراد ALL، ۹۰ درصد و حساسیت آن ۹۳ درصد بدست آمد که نشان‌دهنده عملکرد خوب این دو روش است. یکی از نمونه‌ها که بر اساس یافته‌های بالینی در گروه AML قرار دارد طبق نتایج تمام روش‌های خوشه‌بندی مورد استفاده در گروه ALL قرار گرفت که از نظر بالینی می‌تواند قابل توجه باشد.**

**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 زیرگروه‌هایی از نمونه‌ها را به نحوی متمایز کند که برای انطباق آن با یافته‌های بالینی، پژوهش‌های آزمایشگاهی یا بالینی جدیدی لازم باشد.**

**واژگان کلیدی: بیوانفورماتیک، ریزآرایهDNA، بیان ژن، خوشه بندی، سرطان خون**

## مقدمه

سرطان‌ها بدون شک امروزه یکی از مهم‌ترین عوامل مرگ و میر هستند. به‌طوری‌که پس از بیماری‌های قلبی و تصادفات به عنوان سومین عاملی شناخته می‌شوند که سال‌های زیادی از عمر انسان را تلف می‌کنند [۱]. به همین دلیل حجم وسیعی از مطالعات پزشکی به سمت شناسایی، درمان و پیش‌گیری از

* نویسنده مسئول، تلفن: ۰۹۱۲۱۴۸۳۶۸۷ نمابر: ۰۲۱ ۲۲۷۲۱۱۵۰ Email: alavimajd@gmail.com

تاریخ دریافت: ۱۳۸۶/۹/۶ تاریخ پذیرش: ۱۳۸۷/۴/۵



[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]

آن‌ها هدایت می‌شود. شناسایی عوامل ایجادکننده سرطان و روش‌های مختلف تشخیص و درمان آن‌ها نیز بخش مهمی از این مطالعات را تشکیل می‌دهند.

سرطان خون حاد یکی از این سرطان‌هایی است که در صورت عدم شناسایی بهنگام، بیمار را به سرعت از پای در می‌آورد. به منظور درمان سرطان خون حاد می‌بایست ابتدا این بیماری را در دسته‌ها و گروه‌های همگن طبقه‌بندی کرد، با پیش‌رفت تحقیقات ژنتیکی و کشف این موضوع که جهش‌ها و نقایص ژنتیکی از عمده‌ترین دلایل ایجاد بیماری هستند، ایده یافتن گروه‌های همگن سرطان‌ها بر اساس رفتار ژنتیکی‌شان در ذهن محققان ایجاد شد تا با خوشه‌بندی سرطان خون بر اساس عوامل ژنتیکی در زیرگروه‌های همگن فرایند تشخیص و درمان آن‌ها را تسریع بخشند.

بیوانفورماتیک (Bioinformatics) و فناوری‌های جدید روش‌های بیولوژیکی سنتی را دچار تحول نموده‌اند. تلفیق ابزارهای محاسباتی و دستگاه‌های پیچیده مهندسی زمینه را برای کشف حیطه‌های خاصی همچون ژنتیک که تاکنون ناشناخته مانده‌اند بیش از پیش فراهم نموده است. حاصل این تلفیق را می‌توان ظهور فناوری نوین ریزآرایه (Microarray) DNA در سال ۱۹۹۵ دانست [۲]. ریزآرایه به پژوهش‌گران امکان بررسی هم‌زمان بیان هزاران ژن در حداقل زمان ممکن را می‌دهد. از سال ۱۹۹۸ خوشه‌بندی داده‌های بیان ژنی شروع گردیده است [۳]. در میان روش‌های مختلف خوشه‌بندی، روش خوشه‌بندی سلسله مراتبی (Hierarchical) به دلیل نحوه نمایش و قراردادن تک‌تک عناصر در خوشه‌ها به عنوان یکی از رایج‌ترین روش‌های خوشه‌بندی داده‌ها در فناوری ریزآرایه معرفی می‌گردد [۴]. مطالعات زیادی برای بررسی بیان ژنی داده‌های سرطان خون طراحی گردیده است، که با توجه به آنها محققان توانستند برخی از خوشه‌بندی‌ها موجود را اصلاح و خوشه‌های جدیدی را معرفی نمایند و بقای بیماران در خوشه‌های مختلف را با استفاده از روش‌های آماری بررسی کنند [۵ و۶]. ریزآرایه ابزاری برای اندازه‌گیری و کسب اطلاعات از بیان ژن‌هاست. هر توالی ژنی شناخته شده مورد علاقه 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 (Complementary DNA Spotted)

۲ - آرایه الیگونوکلئوتید (Oligonucleotide array) که به اختصار الیگو گفته می‌شود [۷].

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 علامت‌گذاری می‌شوند. سپس با هدف تحریک رنگ‌های فلورسنت در دو طول موج مختلف آرایه بوسیله 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 چیپی خواهد بود که حاوی هزاران لکه رنگی با رنگ‌های بسیار متنوع است که از ترکیب دو رنگ قرمز و سبز حاصل شده است. یک اندازه بیان ژنی می‌تواند لگاریتم نسبت شدت رنگ قرمز به سبز باشد [۸-۱۰].

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 کاملی (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 غیر جفت (MisMatch) یا MM می‌گویند. یک اندازه از بیان ژنی به صورت متوسط شدت اختلافات در این ۱۶ تا ۲۰ حالت می‌باشد [۷ و ۱۱]. یک برآورد جایگزین برای بیان ژن‌هایی، که از روش الیگواری بدست می‌آیند به



[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]

صورت لگاریتم مقادیر PM از ۱/۲ لگاریتم مقادیر MM و میانگین آن‌ها بر روی ۱۶ تا ۲۰ جفت بیان می‌شود و مشخص گردیده که این برآورد در پیدا کردن ژن‌هایی که به طور مختلفی بیان شده‌اند از روش میانگین دقیق‌تر عمل می‌کند [۱۲].

اغلب داده‌های حاصل از این دو روش در ماتریسی به عنوان ماتریس بیان ژنی (Gene Expression) ذخیره می‌شود که سطرهای آن ژن‌ها و ستون‌های آن افراد نمونه می‌باشند. 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].

هدف مشترک در مطالعات بیان ژنی که بر اساس آن 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 زیر نمونه‌های 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 که بیان آنها دچار تغییر شده است [۷]. هدف این تحقیق جواب به قسمت اول با استفاده از روش خوشه‌بندی است. روش‌های مختلف خوشه‌بندی با توجه به گروه‌بندی واقعی نمونه‌ها (ALL، AML) مقایسه می‌شوند.

## مواد و روش‌ها

از آنجایی که ریزآرایه DNA و روش‌های تحلیل داده‌های بدست 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 نظیر Gen Bank استفاده کنند [۱۴].

در این تحقیق از داده‌های سرطان خون که توسط گلوب و همکاران (Golub) در سال ۱۹۹۹ انتشار یافته استفاده شد [۱۵]. نمونه‌های سرطان خون شامل ۲۴ نمونه مغز استخوان و ۱۰ نمونه خون می‌باشد که همگی در زمان تشخیص سرطان خون گرفته شده‌اند. ۲۰ نمونه از بیماران با سرطان خون حاد لنفوئیدی (ALL) و ۱۴ نمونه از بیماران با سرطان خون حاد میلوئیدی (AML) می‌باشند، که به‌صورت نمونه‌گیری مبتنی بر هدف و غیرتصادفی انتخاب شده‌اند و بر اساس روش آرایه الیگونوکلئوتید بیان ژن‌ها بدست آمده است. این داده‌ها از طریق اینترنت در اختیار عموم قرار دارد [۱۶].

در این داده‌ها هم نظیر بیش‌تر داده‌های مطالعات بیان ژنی، داده‌ها چوله و دارای نقاط پرت بودند، لذا لازم بود ابتدا یک پیش پردازش بر روی داده‌ها صورت گیرد.

با توجه به توصیه‌های گلوب و همکاران (۱۹۹۹) موارد ذیل برای پیش پردازش داده‌ها در نظر گرفته شد:

۱) انتخاب حد آستانه برای داده‌ها: حداقل مقدار هر داده ۱۰۰ و حداکثر مقدار ۱۶۰۰۰ باشد، یعنی داده‌هایی که مقدار بیان ژنی آن‌ها کمتر از ۱۰۰ بود را ۱۰۰ منظور گردید و داده‌هایی که بیشتر از ۱۶۰۰۰ بودند را ۱۶۰۰۰ در نظر گرفته شد.

۲) فیلتر کردن: خارج کردن داده‌هایی که $\max/\min \le 5$ و $\max - \min \le 500$ بودند. منظور از max و min به ترتیب حداکثر و حداقل سطوح بیان یک ژن خاص در کل ۳۴ نمونه می‌باشد.

۳) انجام تبدیل لگاریتمی بر روی داده‌ها.

۴) استاندارد کردن: تبدیل نرمال استاندارد بر روی داده‌ها انجام شد در نتیجه برای هر نمونه سطوح بیان ژنی دارای میانگین صفر و واریانس یک شد [۱۵].

نمایش گرافیکی نتایج اجازه می‌دهد تا الگوی داده‌ها و گروه‌بندی نمونه‌ها را بدون مشخص کردن تعداد گروه‌های مورد نظر تشخیص دهیم. در ابتدا همه نمونه‌ها به صورت یک بردار $p$ بعدی بیان ژنی ($p$ تعداد ژن‌ها پس از پیش پردازش) نشان داده می‌شوند. هدف مقیاس‌بندی چند بعدی کاهش



[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]

داده‌ها به یک بردار $q$ بعدی ($q<p$) است و نشان دادن نمونه‌ها در این بعد کوچک‌تر است تا الگوی نهفته در داده‌ها تشخیص داده شود.

خوشه‌بندی نمونه‌ها با استفاده از نمودارهای مقیاس‌بندی چند بعدی صورت می‌گیرد. برای تعیین بهترین بعد می‌توان از نمودار تنش در مقابل ابعاد استفاده کرد. هر بعدی که کم‌ترین مقدار تنش را داشته باشد بهترین است [۱۸].

روش خوشه‌بندی سلسله مراتبی با یک سری ادغام‌ها یا تقسیمات متوالی از آزمودنی‌های واقع در تحلیل 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 که بر اساس ماتریس داده‌ها به دست آمده خوشه‌بندی انجام گردید.

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

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \xrightarrow{transformed} D_{n\times n}$$

معیار شباهتی که در این پژوهش استفاده گردید ضریب همبستگی پیرسون (Pearson's Correlation Coefficient) است که به صورت:

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

و معیار اختلاف $d_{ij}=1-\rho_{ij}$ در نظر گرفته شد.

این تکنیک خوشه‌بندی خود دو نوع است: تجمعی (Agglomerative) و تقسیمی .(Divisive) نتایج هر دو این روش‌ها قطعی و برگشت ناپذیر است. بدین مفهوم که اگر در روش تجمعی 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. در این تحقیق از روش خوشه‌بندی سلسله مراتبی تجمعی ادغام بر اساس مراکز خوشه‌ای(میانگین) و نیز روش خوشه‌بندی سلسله مراتبی تقسیمی استفاده گردید.

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، ارتفاع هر یک از شاخه‌ها بیان‌گر آن است که دو خوشه مورد نظر در چه نقطه‌ای با هم ادغام شده‌اند [۱۷]. برای مقایسه نمودارهای درختی از معیار فاصله کوفنتیک (Cophenetic Distances) که نشان‌دهنده فواصل واقعی خوشه‌هاست استفاده گردید، هر نموداری که معیار فاصله کوفنتیک آن بالاتر باشد خوشه‌بندی داده‌ها را بهتر نشان می‌دهد [۱۹].

روش‌های خوشه‌بندی غیر سلسله مراتبی (Non Hierarchical) برای دسته بندی کردن اقلام بجای متغیرها به مجموعه‌ای از $k$ خوشه طراحی شده است. یکی از روش‌های غیر سلسله مراتبی که در این تحقیق از آن استفاده کرده‌ایم روش افراز کردن اطراف میدوئید (Partitioning Around Medoids یا $k$ – میدوئید است. الگوریتمی که در روش $PAM$ استفاده می‌شود بر پایه جستجو برای $k$ عنصر نماینده در میان عناصر یک مجموعه داده است. این عناصر باید نماینده جنبه‌های ساختار داده‌ها باشند. در الگوریتم $PAM$ این عناصر به عنوان میدوئیدهای خوشه‌ها نامیده می‌شوند، که عناصری از خوشه است که با توجه به آن متوسط اختلاف همه عناصر از آنها مینیمم باشد. بعد از اینکه یک مجموعه $k$ عنصری از عناصر نماینده را پیدا کردیم، $k$ خوشه بر اساس اختصاص دادن عناصر نزدیک به نماینده شکل می‌گیرد [۱۸].

## نتایج

برای هر کدام از نمونه‌های مورد استفاده مقدار بیان ژنی ۷۱۲۹ ژن جمع‌آوری گردیده است که پس از پیش‌پردازش اولیه ۲۹۱۷ ژن باقی ماندند. در اولین مرحله ماتریس فاصله با ابعاد ۳۴×۳۴ که هر عضو آن برابر یک منهای ضریب

[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]

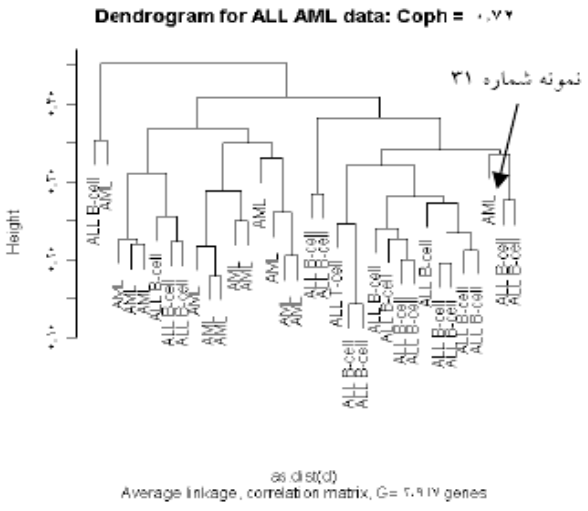
همبستگی پیرسون بین بیماران بر مبنای ۲۹۱۷ ژن باقیمانده بود تشکیل گردید. برای تعیین بهترین بعد از نمودار تنش در مقابل ابعاد استفاده کردیم (برای اختصار نمودار تنش بیان نشده است). بین یک و دو بعد اختلاف زیاد است ولی از دو بعد به بالا با افزایش بعد شاخص تنش زیاد تغییر نمی‌کند و از آنجایی که نمایش ابعاد بالاتر مشکل می‌باشد لذا خوشه‌بندی را با توجه به مقیاس‌بندی سه بعدی انجام دادیم، که دو خوشه را می‌توان بین نمونه‌ها تشخیص داد. نتیجه انتساب به دو خوشه در جدول ۱ آمده است. با توجه به این جدول ویژگی روش مقیاس‌بندی سه بعدی در تشخیص افراد ALL، ۹۵ درصد و حساسیت آن ۹۳ درصد بدست آمد.

با انجام خوشه‌بندی سلسله مراتبی تجمعی معیار فاصله کوفنتیک برای ادغام برحسب میانگین ۰/۷۲، ادغام کامل ۰/۶۲ و ادغام منفرد ۰/۶۲ است. در شکل ۱ نتیجه خوشه‌بندی سلسله مراتبی تجمعی بر اساس ادغام بر حسب میانگین نشان داده شده است. با توجه به این شکل نمونه‌ها را می‌توان در دو خوشه دسته‌بندی کرد. با توجه به جدول ۱ ویژگی روش خوشه‌بندی سلسله مراتبی در تشخیص افراد ALL، ۹۵ درصد بدست آمد. با انجام خوشه‌بندی سلسله مراتبی تقسیمی معیار فاصله کوفنتیک ۰/۶۸ بدست آمد. در شکل ۲ نتیجه خوشه‌بندی با این روش نشان داده شده است. با توجه به جدول ۱ ویژگی روش خوشه‌بندی سلسله مراتبی تقسیمی در تشخیص افراد ALL، ۷۵ درصد و حساسیت آن ۹۲ درصد به‌دست آمد.

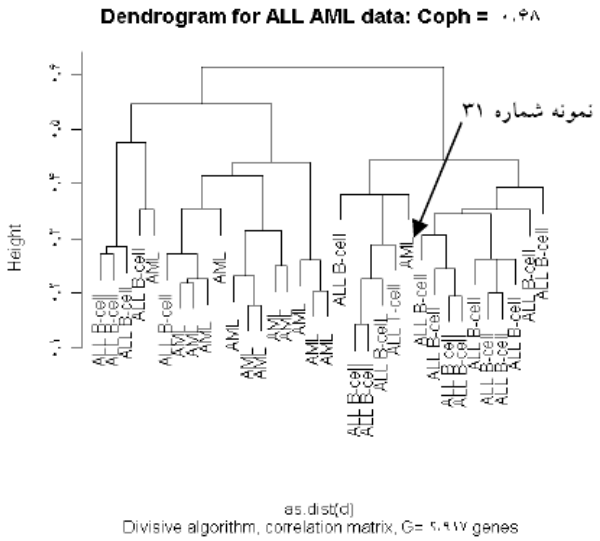
در شکل ۳ روش افراز کردن اطراف میدوئید در اختصاص به دو خوشه نشان داده شده است. با توجه به جدول ۱ ویژگی روش افراز کردن اطراف میدوئید در تشخیص افراد ALL، ۹۰ درصد و حساسیت آن ۹۳ درصد به‌دست آمد.

جدول۱. انتساب بیماران با روش مقیاس‌بندی سه بعدی، سلسله مراتبی تجمعی، سلسله مراتبی تقسیمی و افراز کردن اطراف میدوئید به دو خوشه

| خوشه بندی / سرطان | مقیاس‌بندی سه بعدی | | سلسله مراتبی تجمعی | | سلسله مراتبی تقسیمی | | افراز کردن اطراف میدوئید | |
|---|---|---|---|---|---|---|---|---|
| | خوشه۱ | خوشه۲ | خوشه۱ | خوشه۲ | خوشه۱ | خوشه۲ | خوشه۱ | خوشه۲ |
| ALL | ۱۹ | ۱ | ۱۹ | ۱ | ۵ | ۱۵ | ۱۸ | ۲ |
| AML | ۱ | ۱۳ | ۱۳ | ۱ | ۱۳ | ۱ | ۱ | ۱۳ |
| حساسیت | ٪۹۳ | | ٪۷ | | ٪۹۳ | | ٪۹۳ | |
| ویژگی | ٪۹۵ | | ٪۹۵ | | ٪۷۵ | | ٪۹۰ | |



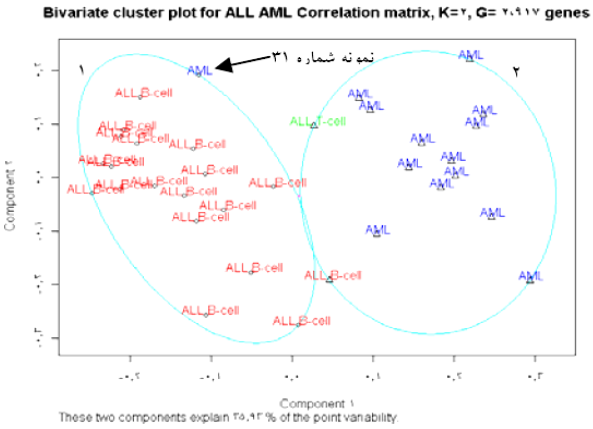

شکل ۱. نمودار درختی روش خوشه‌بندی سلسله مراتبی تجمعی بر اساس مراکز خوشه‌ای (میانگین)




شکل ۲. نمودار درختی خوشه‌بندی سلسله مراتبی تقسیمی



[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]




شکل۳. نمایش نمونه‌ها در دو خوشه در روش افراز کردن اطراف میدوئید

## بحث و نتیجه‌گیری

در این مقاله داده‌های بیان ژنی سرطان خون با استفاده از روش‌های آماری (مقیاس‌بندی چند بعدی، خوشه‌بندی سلسله مراتبی و غیر سلسله مراتبی) خوشه‌بندی گردید. با توجه به نتایج روش مقیاس‌بندی در انتساب نمونه‌ها به دو خوشه توانایی کامل دارد زیرا از ۲۰ نمونه ALL ۱۹نمونه و از ۱۴ نمونه AML ۱۳ نمونه را به درستی اختصاص داده است. با مقایسه بین سه معیار فاصله کوفنتیک مشخص شد که روش خوشه‌بندی سلسله مراتبی تجمعی ادغام میانگین مناسب‌تر از دیگر روش‌های خوشه‌بندی سلسله مراتبی تجمعی است. با در نظر گرفتن شکل دیده می‌شود هر اندازه که تعداد خوشه‌ها بیش‌تر شود روش خوشه‌بندی سلسله مراتبی تجمعی ادغام میانگین زیرگروه‌های مجزایی از ALL و AML را ارایه می‌دهد که برای بررسی‌های بیش‌تر شناسایی چنین زیرگروه‌هایی ضروری است. معیار کوفنتیک روش خوشه‌بندی سلسله مراتبی تقسیمی ۰/۶۸ می‌باشد که تنها از روش خوشه‌بندی سلسله مراتبی ادغام میانگین کمتر است. این روش در اختصاص نمونه‌ها به دو خوشه از روش ادغام میانگین نیز بهتر عمل کرد.

نمونه شماره ۳۱ (در نمودارها با پیکان مشخص گردیده) بر اساس تشخیص‌های بالینی در گروه AML قرار گرفته است اما نتایج خوشه‌بندی نشان می‌دهدکه نمونه شماره ۳۱ در گروه ALL قرار می‌گیرد، بنابراین موارد عدم انطباق به این شکل می‌تواند موضوع بررسی‌های بالینی بیشتر و دقیق‌تر باشد.

از بین روش‌هایی که بررسی گردید روش مقیاس‌بندی و روش خوشه‌بندی غیر سلسله مراتبی افراز کردن اطراف میدوئید با توجه به مقادیر حساسیت و ویژگی و انطباق بیش‌تر با گروه‌بندی واقعی افراد نمونه نتایج بهتری را نسبت به روش‌های سلسله مراتبی ارایه دادند. از آنجایی 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 ساده‌تر هستند و می‌توان آنها را به مجموعه داده‌های بسیار بزرگ‌تری نسبت به روش‌های سلسله مراتبی به‌کار برد.

همانطور که در این مقاله دیده شد بکارگیری روش‌های آماری خوشه‌بندی در مقایسه با یافته‌های بالینی می‌تواند دیدگاه‌های جدیدی در تشخیص و افتراق بیماری به روی محققان بگشاید. قطعاً با توجه به اینکه توالی کامل ژنوم انسانی در دسترس قرار گرفته است، استفاده از روش‌های آماری در بکارگیری داده‌های ریزآرایه می‌تواند موضوع پژوهش‌های بعدی محققان باشد.

## تشکر و قدردانی

این مقاله حاصل طرح تحقیقاتی است که اعتبار آن توسط دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی تامین شده است که در اینجا از معاونت محترم آموزشی پژوهشی دانشکده سپاس‌گزاری می‌شود.

## منابع

[1] Haskell C.M, Bereck J. Cancer Treatment. Sounder co, 2001. 5th.Edition PP10-21

[2] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270: 467-470.

[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: 14863–14868.

[4] Gershon D. Microarray technology: an array of opportunities. Nature 2002; 416: 885-891.

[5] Bullinger L, Dhner K, Bair E, Frhling S, Schlenk RF, Tibshirani R, and et al. Use of gene expression profiling to

[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]

identify prognostic subclasses in adult acute myeloid leukemia. The New England Journal of Medicine 2004; 350: 1605-1616.

[6] Valk PJM, Verhaak RGW, Beijen M A, Erpelinck CAJ, Barjesteh S, Waalwijk V, et al. Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. New Eng J. Med 2004; 350: 1617-1628.

[7] Satagopan JM, Panageas KS, Tutorial in biostatistics a statistical perspective on gene expression data analysis. Statist. Med. 2003; 22:481–499.

[8] Eisen MB, Brown PO. DNA arrays for analysis of gene expression. Methods Enzymolo 1999; 303: 179 –205.

[9] Amaratunga D, Cabrera J. Exploration and Analysis of DNA Microarray and Protein Array Data. Wiley & Sons, Ltd, 2004; 1.Edition PP 8-37

[10] Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed Optics 1997; 2: 364 –374.

[11] Affymetrix Microarray Suite User Guide. Version 4.0. 2000; Appendix A2, A3.

[12] Efron B, Tibshirani R, Goss V, Chu G. Microarrays and their use in a comparative experiment. Technical report 213, Department of Statistics, Stanford University, 2002. Available from: URL: http://www-stat.stanford.edu/tibs/research. html

[13] The R Project for Statistical Computing Available from: URL: http://www.r-project.org

[14] National Center for Biotechnology Information Available from: URL: http://www. ncbi.nlm.nih.gov.

[15] 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: 531–537.

[16] Available from: URL: http://www.broad.mit.edu/MPR

[18] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to cluster Analysis. Wiley: New York, 2005; 1.Edition : PP 68-279

[19] Sneath P.H. and Sokal R.R. The principles and practice of numerical classification. Numerical Taxonomy. W. H. Freeman, San Francisco, 1973: p.278 ff.

[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]

# Gene expression data clustering and it's application in differential analysis of leukemia

M. Vahedi (M.Sc) [*1], H. Alavi Majd (Ph.D) [2], Y. Mehrabi (Ph.D) [2], B. Naghavi (M.Sc)[2]

*1 - Research Center for Gastroenterology and Liver Diseases, Shaheed Beheshti Medical University, Tehran, Iran.*

*2 - Shaheed Beheshti Medical University, Tehran, Iran.*

***Introduction:*** DNA microarray technique is one of the most important categories in bioinformatics, which allows the possibility of monitoring thousands of expressed genes has been resulted in creating giant data bases of gene expression data, recently. Statistical analysis of such databases included normalization, clustering, classification and etc.

***Materials and Methods:*** Golub et al (1999) collected data bases of leukemia based on the method of oligonucleotide. The data is on the internet. In this paper, we analyzed gene expression data. It was clustered by several methods including multi-dimensional scaling, hierarchical and non-hierarchical clustering. Data set included 20 Acute Lymphoblastic Leukemia (ALL) patients and 14 Acute Myeloid Leukemia (AML) patients. The results of tow methods of clustering were compared with regard to real grouping (ALL & AML). R software was used for data analysis.

***Results:*** Specificity and sensitivity of divisive hierarchical clustering in diagnosing of ALL patients were 75% and 92%, respectively. Specificity and sensitivity of partitioning around medoids in diagnosing of ALL patients were 90% and 93%, respectively. These results showed a well accomplishment of both methods of clustering. It is considerable that, due to clustering methods results, one of the samples was placed in ALL groups, which was in AML group in clinical test.

***Conclusion:*** With regard to concordance of the results with real grouping of data, therefore we can use these methods in the cases where we don't have accurate information of real grouping of data. Moreover, Results of clustering might distinct subgroups of data in such a way that would be necessary for concordance with clinical outcomes, laboratory results and so on.

***Key words:*** Bioinformatics, DNA microarray, Gene expression, Clustering, Leukemia

* Corresponding author: Fax: +98 021 22721150; Tel: 09121483687
alavimajd@gmail.com

[ Downloaded from koomeshjournal.semums.ac.ir on 2024-07-31 ]