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 ... شتی، مرکز تحقیقات بیماری‌های گوارش و کبد

۲ - دانشگاه علوم پزشکی شهید ... شتی، دانشکده پیرایزشکی

۳ - دانشگاه علوم پزشکی شهید ... شتی، دانشکده ... داشت

۴ - دانشگاه علوم پزشکی شهید ... شتی، 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 شك امروزه يكي از مهم‌ترين عوامل مرگ و مير هستند. به‌طوري‌كه پس از بيماري‌هاي قلبي و تصادفات به عنوان سومين عاملي شناخته مي‌شوند كه سال‌هاي زيادي از عمر انسان را تلف مي‌كنند [۱]. به همين دليل حجم وسيعي از مطالعات پزشكي به سمت شناسايي، درمان و پيش‌گيري از

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

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

آن‌ها هدايت مي‌شود. شناسايي عوامل ايجادكننده سرطان و روش‌هاي مختلف تشخيص و درمان آن‌ها نيز بخش مهمي از اين مطالعات را تشكيل مي‌دهند.

سرطان خون حاد يكي از اين سرطان‌هايي است كه در صورت عدم شناسايي نگام، بيمار را به سرعت از پاي در مي‌آورد. به منظور درمان سرطان خون حاد مي‌بايست ابتدا اين بيماري را در دسته‌ها و گروه‌هاي همگن طبقه‌بندي كرد، با پيش‌رفت تحقيقات ژنتيكي و كشف اين موضوع كه جهش‌ها و نقايص ژنتيكي از عمده‌ترين دلايل ايجاد بيماري هستند، ايده يافتن گروه‌هاي همگن سرطان‌ها بر اساس رفتار ژنتيكي‌شان در ذهن محققان ايجاد شد تا با خوشه‌بندي سرطان خون بر اساس عوامل ژنتيكي در زيرگروه‌هاي همگن فرايند تشخيص و درمان آن‌ها را تسريع بخشند.

بيوانفورماتيك (Bioinformatics) و فناوري‌هاي جديد روش‌هاي بيولوژيكي سنتي را دچار تحول نموده‌اند. تلفيق ابزارهاي محاسباتي و دستگاه‌هاي پيچيده مهندسي زمينه را براي كشف حيطه‌هاي خاصي همچون ژنتيك كه تاكنون ناشناخته مانده‌اند بيش از پيش فراهم نموده است. حاصل اين تلفيق را مي‌توان ظهور فناوري نوين ريزآرايه (Microarray) DNA در سال ۱۹۹۵ دانست [۲]. ريزآرايه به پژوهش‌گران امكان بررسي هم‌زمان بيان هزاران ژن در حداقل زمان ممكن را مي‌دهد. از سال ۱۹۹۸ خوشه‌بندي داده‌هاي بيان ژني شروع گرديده است [۳]. در ميان روش‌هاي مختلف خوشه‌بندي، روش خوشه‌بندي سلسله مراتبي (Hierarchical) به دليل نحوه نمايش و قراردادن تك‌تك عناصر در خوشه‌ها به عنوان يكي از رايج‌ترين روش‌هاي خوشه‌بندي داده‌ها در فناوري ريزآرايه معرفي مي‌گردد [۴]. مطالعات زيادي براي بررسي بيان ژني داده‌هاي سرطان خون طراحي گرديده است، كه با توجه به آ ا محققان توانستند برخي از خوشه‌بندي‌ها موجود را اصلاح و خوشه‌هاي جديدي را معرفي نمايند و بقاي بيماران در خوشه‌هاي مختلف را با استفاده از روش‌هاي آماري بررسي كنند [۵ و۶]. ريزآرايه ابزاري براي اندازه‌گيري و كسب اطلاعات از بيان ژن‌هاست. هر توالي ژني شناخته شده مورد علاقه به عنوان يك پروب (Prob) روي يك آرايه (Array) شيشه‌اي يا نايلوني چاپ مي‌شود mRNA . از بافت يا نمونه خون با رنگ‌هاي فلورسنت علامت‌گذاري مي‌شود و پروب‌ها بر روي يك آرايه هيبريد مي‌شوند. دو نوع آرايه بيش‌ترين كاربرد را دارند:

۱- آرايه‌هاي بر پايه DNA مكمل (Complementary DNA Spotted)

۲ - آرايه اليگونوكلئوتيد (Oligonucleotide array) كه به اختصار اليگو گفته مي‌شود [۷].

در روش آرايه cDNA هر ژن با يك رشته طولاني (بين ۲۰۰ تا ۵۰۰ پايه) نشان داده مي‌شود cDNA از دو نمونه متفاوت بدست مي‌آيد، يكي از نمونه‌ي مورد آزمون و ديگري نمونه‌ي مرجع كه بر روي يك آرايه هيبريد (مخلوط) مي‌شوند. نمونه آزمون با رنگ فلورسنت قرمز و نمونه مرجع با رنگ سبز علامت‌گذاري مي‌شوند. سپس با هدف تحريك رنگ‌هاي فلورسنت در دو طول موج مختلف آرايه بوسيله اشعه ليزر اسكن مي‌شود. از هر كدام اين رنگ‌ها يك تصوير بوجود مي‌آيد كه اين تصاوير در كامپيوتر بر روي هم قرار داده مي‌شود كه حاصل اين كار چيپي خواهد بود كه حاوي هزاران لكه رنگي با رنگ‌هاي بسيار متنوع است كه از تركيب دو رنگ قرمز و سبز حاصل شده است. يك اندازه بيان ژني مي‌تواند لگاريتم نسبت شدت رنگ قرمز به سبز باشد [۸-۱۰].

در روش آرايه اليگونوكلئوتيد هر ژن به ۱۶ الي ۲۰ حالت نشان داده مي‌شود كه هر كدام خود توالي كوتاهي از نوكلئوتيدها هستند و يك جفت كاملي (Perfect Match) يا PM از يك قطعه ژن مي‌باشد، در مقابل اين ۲۰ اليگونوكلئوتيد، ۲۰ اليگونوكلئوتيد ديگر وجود دارد كه به جز در باز مركزي توالي آ ا با هم برابر است، كه به اين نوكلئوتيدها غير جفت (MisMatch) يا MM مي‌گويند. يك اندازه از بيان ژني به صورت متوسط شدت اختلافات در اين ۱۶ تا ۲۰ حالت مي‌باشد [۷ و ۱۱]. يك برآورد جايگزين براي بيان ژن‌هايي، كه از روش اليگواري بدست مي‌آيند به صورت

لگاریتم مقادیر PM از ۱/۲ لگاریتم مقادیر MM و میانگین آن‌ها بر روي ۱۶ تا ۲۰ جفت بیان مي‌شود و مشخص گردیده که این برآورد در پیدا کردن ژن‌هایي که به طور مختلفي بیان شده‌اند از روش میانگین دقیق‌تر عمل مي‌کند [۱۲].

اغلب داده‌هاي حاصل از این دو روش در ماتریسي به عنوان ماتریس بیان ژني (Gene Expression) ذخیره مي‌شود که سطرهاي آن ژن‌ها و ستون‌هاي آن افراد نمونه مي‌باشند. براي بررسي داده‌هاي ریزآرایه DNA مي‌توان از نرم‌افزارهاي آماري نظیر SAS ، S-plus، STATA و R استفاده کرد . R به علت توانایي بالا در کار کردن با داده‌هاي حجیم رواج بیش‌تري دارد [۱۳].

هدف مشترك در مطالعات بیان ژني که بر اساس آن از روش‌هاي تحلیل آماري متفاوتي استفاده مي‌شود اغلب در یکي از این موارد قرار دارد: ۱- کشف طبقات (Class discovery) ۲- تشخیص ژني (Gene identification) ۳ - پیش‌بیني طبقات (Class prediction). کشف طبقات شناسایي زیر نمونه‌هاي نامعلوم است در حالي که روش‌هاي پیش‌بیني طبقات براي طبقه‌بندي نمونه مستقل جدید بر اساس یك طرح پیش‌بیني مورد استفاده قرار مي‌گیرند. تشخیص ژني مربوط به شناسایي ژن‌هایي که بیان آ ا دچار تغییر شده است [۷]. هدف این تحقیق جواب به قسمت اول با استفاده از روش خوشه‌بندي است. روش‌هاي مختلف خوشه‌بندي با توجه به گروه‌بندي واقعي نمونه‌ها (ALL، AML) مقایسه مي‌شوند.

مواد و روش‌ها

از آنجایي که ریزآرایه DNA و روش‌هاي تحلیل داده‌هاي بدست آمده از آن جدید مي‌باشد، تاکنون در بسیاري از کشورها از جمله ایران کار زیادي براي تولید این گونه داده‌ها صورت نگرفته است، بنابراین اغلب محققان ناچارند که از داده‌هاي بانک‌هاي اطلاعاتي اینترنتي نظیر Gen Bank استفاده کنند [۱٤].

در این تحقیق از داده‌هاي سرطان خون که توسط گلوب و همکاران (Golub) در سال ۱۹۹۹ انتشار یافته استفاده شد [۱۵]. نمونه‌هاي سرطان خون شامل ۲٤ نمونه مغز استخوان و ۱۰ نمونه خون مي‌باشد که همگي در زمان تشخیص سرطان خون گرفته شده‌اند. ۲۰ نمونه از بیماران با سرطان خون حاد لنفوئیدي (ALL) و ۱٤ نمونه از بیماران با سرطان خون حاد میلوئیدي (AML) مي‌باشند، که به‌صورت نمونه‌گیري مبتني بر هدف و غیرتصادفي انتخاب شده‌اند و بر اساس روش آرایه الیگونوکلئوتید بیان ژن‌ها بدست آمده است. این داده‌ها از طریق اینترنت در اختیار عموم قرار دارد [۱٦].

در این داده‌ها هم نظیر بیش‌تر داده‌هاي مطالعات بیان ژني، داده‌ها چوله و داراي نقاط پرت بودند، لذا لازم بود ابتدا یك پیش پردازش بر روي داده‌ها صورت گیرد.

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

۱) انتخاب حد آستانه براي داده‌ها: حداقل مقدار هر داده ۱۰۰ و حداکثر مقدار ۱٦۰۰۰ باشد، یعني داده‌هایي که مقدار بیان ژني آن‌ها کمتر از ۱۰۰ بود را ۱۰۰ منظور گردید و داده‌هایي که بیشتر از ۱٦۰۰۰ بودند را ۱٦۰۰۰ در نظر گرفته شد.

۲) فیلتر کردن: خارج کردن داده‌هایي که $\max/\min \le 5$ و $\max - \min \le 500$ بودند. منظور از max و min به ترتیب حداکثر و حداقل سطوح بیان یك ژن خاص در کل ۳٤ نمونه مي‌باشد.

۳) انجام تبدیل لگاریتمي بر روي داده‌ها.

٤) استاندارد کردن: تبدیل نرمال استاندارد بر روي داده‌ها انجام شد در نتیجه براي هر نمونه سطوح بیان ژني داراي میانگین صفر و واریانس یك شد [۱۵].

نمایش گرافیکي نتایج اجازه مي‌دهد تا الگوي داده‌ها و گروه‌بندي نمونه‌ها را بدون مشخص کردن تعداد گروه‌هاي مورد نظر تشخیص دهیم. در ابتدا همه نمونه‌ها به صورت یک بردار p بعدي بیان ژني (p تعداد ژن‌ها پس از پیش پردازش) نشان داده مي‌شوند. هدف مقیاس‌بندي چند بعدي کاهش داده‌ها به یک بردار q بعدي ($q<p$) است و نشان دادن نمونه‌ها در این

بعد كوچك‌تر است تا الگوي نهفته در داده‌ها تشخيص داده شود.

خوشه‌بندي نمونه‌ها با استفاده از نمودارهاي مقياس‌بندي چند بعدي صورت مي‌گيرد. براي تعيين ترين بعد مي‌توان از نمودار تنش در مقابل ابعاد استفاده كرد. هر بعدي كه كم‌ترين مقدار تنش را داشته باشد ترين است [۱۸].

روش خوشه‌بندي سلسله مراتبي با يك سري ادغام‌ها يا تقسيمات متوالي از آزمودني‌هاي واقع در تحليل همراه است. اساس تعيين خوشه در اين روش محاسبه اندازه شباهت‌ها (Similarity) و يا اختلاف‌ها (Distance) بين هر زوج از عناصر مورد مطالعه است. به اين ترتيب بر اساس ماتريس شباهت يا اختلاف كه بر اساس ماتريس داده‌ها به دست آمده خوشه‌بندي انجام گرديد.

عنصر x_{ik} در ماتريس X نشان دهنده بيان ژن i در فرد k است.

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

معيار شباهتي كه در اين پژوهش استفاده گرديد ضريب همبستگي پيرسون (Pearson's Correlation Coefficient) است كه به صورت:

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

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

اين تكنيك خوشه‌بندي خود دو نوع است: تجمعي (Agglomerative) و تقسيمي .(Divisive) نتايج هر دو اين روش‌ها قطعي و برگشت ناپذير است. بدين مفهوم كه اگر در روش تجمعي دو آزمودني در يك خوشه قرار گرفتند ديگر قابل تفكيك نخواهند بود و يا اگر در روش تقسيم در دو خوشه مجزا قرار گرفتند ديگر امكان اينكه در يك خوشه با هم واقع شوند وجود ندارد. در اين تحقيق از روش خوشه‌بندي سلسله مراتبي تجمعي ادغام بر اساس مراكز خوشه‌اي(ميانگين) و نيز روش خوشه‌بندي سلسله مراتبي تقسيمي استفاده گرديد.

نتايج حاصل از روش خوشه‌بندي سلسله مراتبي در نمودارهاي دندوگرام (Dendrogram) يا نمودار درختي نمايش داده مي‌شود. در دندوگرام‌ها محور عمودي فاصله بين خوشه‌ها را اندازه‌گيري مي‌كند، ارتفاع هر يك از شاخه‌ها بيانگر آن است كه دو خوشه مورد نظر در چه نقطه‌اي با هم ادغام شده‌اند [۱۷]. براي مقايسه نمودارهاي درختي از معيار فاصله كوفنتيك (Cophenetic Distances) كه نشان‌دهنده فواصل واقعي خوشه‌هاست استفاده گرديد، هر نموداري كه معيار فاصله كوفنتيك آن بالاتر باشد خوشه‌بندي داده‌ها را تر نشان مي‌دهد [۱۹].

روش‌هاي خوشه‌بندي غير سلسله مراتبي (Non Hierarchical) براي دسته بندي كردن اقلام بجاي متغيرها به مجموعه‌اي از k خوشه طراحي شده است. يكي از روش‌هاي غير سلسله مراتبي كه در اين تحقيق از آن استفاده كرده‌ايم روش افراز كردن اطراف ميدوئيد (Partitioning Around Medoids يا k - ميدوئيد است. الگوريتمي كه در روش *PAM* استفاده مي‌شود بر پايه جستجو براي k عنصر نماينده در ميان عناصر يک مجموعه داده است. اين عناصر بايد نماينده جنبه‌هاي ساختار داده‌ها باشند. در الگوريتم *PAM* اين عناصر به عنوان ميدوئيدهاي خوشه‌ها ناميده مي‌شوند، كه عناصري از خوشه است كه با توجه به آن متوسط اختلاف همه عناصر از آ ا مينيمم باشد. بعد از اينكه يک مجموعه k عنصري از عناصر نماينده را پيدا كرديم، k خوشه بر اساس اختصاص دادن عناصر نزديک به نماينده شكل مي‌گيرد [۱۸].

نتايج

براي هر كدام از نمونه‌هاي مورد استفاده مقدار بيان ژني ۷۱۲۹ ژن جمع‌آوري گرديده است كه پس از پيش‌پردازش اوليه ۲۹۱۷ ژن باقي ماندند. در اولين مرحله ماتريس فاصله با ابعاد ۳۴×۳۴ كه هر عضو آن برابر يك منهاي ضريب همبستگي پيرسون بين بيماران بر مبناي ۲۹۱۷ ژن باقيمانده

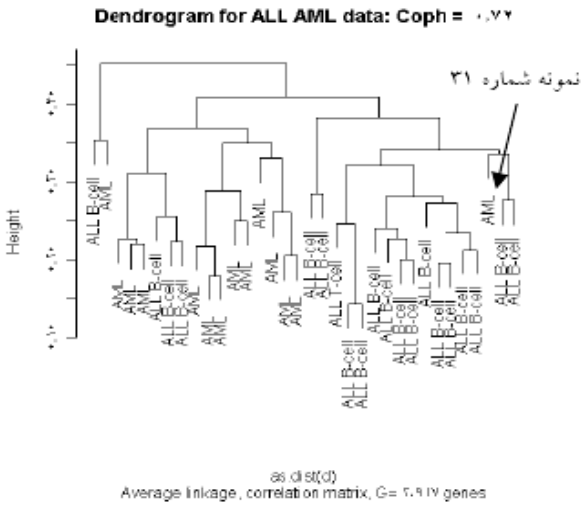
بود تشكيل گرديد. براي تعيين ترين بعد از نمودار تنش در مقابل ابعاد استفاده كرديم (براي اختصار نمودار تنش بيان نشده است). بين يك و دو بعد اختلاف زياد است ولي از دو بعد به بالا با افزايش بعد شاخص تنش زياد تغيير نمي‌كند و از آنجايي كه نمايش ابعاد بالاتر مشكل مي‌باشد لذا خوشه‌بندي را با توجه به مقياس‌بندي سه بعدي انجام داديم، كه دو خوشه را مي‌توان بين نمونه‌ها تشخيص داد. نتيجه انتساب به دو خوشه در جدول ۱ آمده است. با توجه به اين جدول ويژگي روش مقياس‌بندي سه بعدي در تشخيص افراد ALL، ۹۵ درصد و حساسيت آن ۹۳ درصد بدست آمد.

با انجام خوشه‌بندي سلسله مراتبي تجمعي معيار فاصله كوفنتيک براي ادغام برحسب ميانگين ۰/۷۲، ادغام كامل ۰/۶۲ و ادغام منفرد ۰/۶۲ است. در شكل ۱ نتيجه خوشه‌بندي سلسله مراتبي تجمعي بر اساس ادغام بر حسب ميانگين نشان داده شده است. با توجه به اين شكل نمونه‌ها را مي‌توان در دو خوشه دسته‌بندي كرد. با توجه به جدول ۱ ويژگي روش خوشه‌بندي سلسله مراتبي در تشخيص افراد ALL، ۹۵ درصد بدست آمد. با انجام خوشه‌بندي سلسله مراتبي تقسيمي معيار فاصله كوفنتيک ۰/۶۸ بدست آمد. در شكل ۲ نتيجه خوشه‌بندي با اين روش نشان داده شده است. با توجه به جدول ۱ ويژگي روش خوشه‌بندي سلسله مراتبي تقسيمي در تشخيص افراد ALL، ۷۵ درصد و حساسيت آن ۹۲ درصد به‌دست آمد.

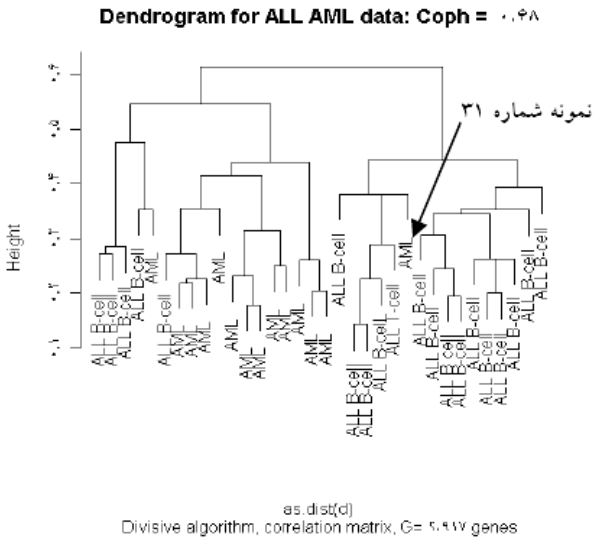
در شكل ۳ روش افراز كردن اطراف ميدوئيد در اختصاص به دو خوشه نشان داده شده است. با توجه به جدول ۱ ويژگي روش افراز كردن اطراف ميدوئيد در تشخيص افراد ALL، ۹۰ درصد و حساسيت آن ۹۳ درصد به‌دست آمد.

جدول۱. انتساب بيماران با روش مقياس‌بندي سه بعدي، سلسله مراتبي تجمعي، سلسله مراتبي تقسيمي و افراز كردن اطراف ميدوئيد به دو خوشه

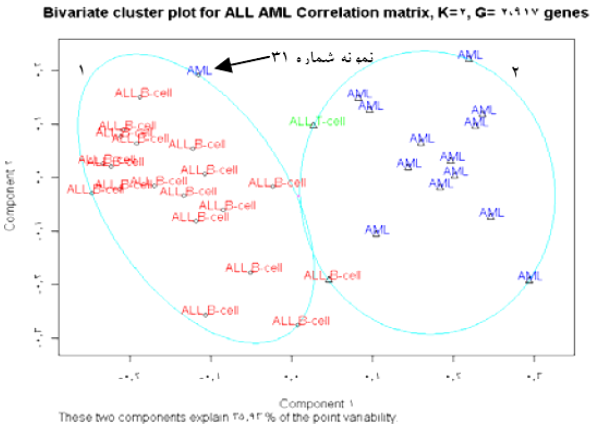
خوشه بندي / سرطان	مقياس‌بندي سه بعدي		سلسله مراتبي تجمعي		سلسله مراتبي تقسيمي		افراز كردن اطراف ميدوئيد	
	خوشه۱	خوشه۲	خوشه۱	خوشه۲	خوشه۱	خوشه۲	خوشه۱	خوشه۲
ALL	۱۹	۱	۱۹	۱	۵	۱۵	۱۸	۲
AML	۱	۱۳	۱۳	۱	۱۳	۱	۱	۱۳
حساسيت	%۹۳		%۷		%۹۳		%۹۳	
ويژگي	%۹۵		%۹۵		%۷۵		%۹۰	



شكل ۱. نمودار درختي روش خوشه‌بندي سلسله مراتبي تجمعي بر اساس مراكز خوشه‌اي (ميانگين)



شكل ۲. نمودار درختي خوشه‌بندي سلسله مراتبي تقسيمي



شكل۳. نمايش نمونه‌ها در دو خوشه در روش افراز كردن اطراف ميدوئيد

بحث و نتيجه‌گيري

در اين مقاله داده‌هاي بيان ژني سرطان خون با استفاده از روش‌هاي آماري (مقياس‌بندي چند بعدي، خوشه‌بندي سلسله مراتبي و غير سلسله مراتبي) خوشه‌بندي گرديد. با توجه به نتايج روش مقياس‌بندي در انتساب نمونه‌ها به دو خوشه توانايي كامل دارد زيرا از ۲۰ نمونه ALL ۱۹نمونه و از ۱۴ نمونه AML ۱۳ نمونه را به درستي اختصاص داده است. با مقايسه بين سه معيار فاصله كوفنتيک مشخص شد كه روش خوشه‌بندي سلسله مراتبي تجمعي ادغام ميانگين مناسب‌تر از ديگر روش‌هاي خوشه‌بندي سلسله مراتبي تجمعي است. با در نظر گرفتن شكل ديده مي‌شود هر اندازه كه تعداد خوشه‌ها بيش‌تر شود روش خوشه‌بندي سلسله مراتبي تجمعي ادغام ميانگين زيرگروه‌هاي مجزايي از ALL و AML را ارايه مي‌دهد كه براي بررسي‌هاي بيش‌تر شناسايي چنين زيرگروه‌هايي ضروري است. معيار كوفنتيك روش خوشه‌بندي سلسله مراتبي تقسيمي ۰/۶۸ مي‌باشد كه تنها از روش خوشه‌بندي سلسله مراتبي ادغام ميانگين كمتر است. اين روش در اختصاص نمونه‌ها به دو خوشه از روش ادغام ميانگين نيز تر عمل كرد.

نمونه شماره ۳۱ (در نمودارها با پيكان مشخص گرديده) بر اساس تشخيص‌هاي باليني در گروه AML قرار گرفته است اما نتايج خوشه‌بندي نشان مي‌دهدكه نمونه شماره ۳۱ در گروه ALL قرار مي‌گيرد، بنابراين موارد عدم انطباق به اين شكل مي‌تواند موضوع بررسي‌هاي باليني بيشتر و دقيق‌تر باشد.

از بين روش‌هايي كه بررسي گرديد روش مقياس‌بندي و روش خوشه‌بندي غير سلسله مراتبي افراز كردن اطراف ميدوئيد با توجه به مقادير حساسيت و ويژگي و انطباق بيش‌تر با گروه‌بندي واقعي افراد نمونه نتايج تري را نسبت به روش‌هاي سلسله مراتبي ارايه دادند. از آنجايي كه در روش خوشه‌بندي غير سلسله مراتبي لازم نيست ماتريس فواصل (مشا ت‌ها) تعيين شود و داده‌هاي اصلي در طول اجرا ذخيره شوند، لذا روش‌هاي غير سلسله مراتبي ساده‌تر هستند و مي‌توان آ ا را به مجموعه داده‌هاي بسيار بزرگ‌تري نسبت به روش‌هاي سلسله مراتبي به‌كار برد.

همانطور كه در اين مقاله ديده شد بكارگيري روش‌هاي آماري خوشه‌بندي در مقايسه با يافته‌هاي باليني مي‌تواند ديدگاه‌هاي جديدي در تشخيص و افتراق بيماري به روي محققان بگشايد. قطعاً با توجه به اينكه توالي كامل ژنوم انساني در دسترس قرار گرفته است، استفاده از روش‌هاي آماري در بكارگيري داده‌هاي ريزآرايه مي‌تواند موضوع پژوهش‌هاي بعدي محققان باشد.

تشكر و قدردانی

اين مقاله حاصل طرح تحقيقاتي است كه اعتبار آن توسط دانشكده پيراپزشكي دانشگاه علوم پزشكي شهيد شتي تامين شده است كه در اينجا از معاونت محترم آموزشي پژوهشي دانشكده سپاس‌گزاري مي‌شود.

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 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.