دانشور
پزشکی

طبقه‌بندی لوکمیای حاد بر مبنای داده‌های بیان ژنی به‌دست آمده از DNA میکرواری با استفاده‌از روش‌های آماری خوشه‌بندی در نرم‌افزار Bioconductor

نویسندگان: دکتر کرامت‌اله‌نوری‌جلیانی[1]، نامعلی آزادی[1]، دکتر حجت‌زراعتی[1] و دکتر سهراب نجفی‌پور[2]

۱- گروه اپیدمیولوژی و آمار زیستی دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران
۲- دانشکده علوم پزشکی و خدمات بهداشتی درمانی فسا

چکیده

سابقه و اهداف: اهمیت روزافزون نقش علل ژنتیکی در ایجاد بیماری‌ها و ظهور روش‌های بیوتکنولوژیکی جدید، باعث شده است که در سال‌های اخیر در تحقیقات مربوط به سبب شناختی بیماری‌ها مطالعات مولکولی اهمیت ویژه‌ای پیدا کنند. DNA میکرواری به‌عنوان یک روش بیوتکنولوژی نوین جهت مطالعه بیان ژن‌ها، در تحقیقات پزشکی کاربردی وسیع یافته است و به نحو گسترده‌ای جهت درک ساختار ژنتیکی نهفته بیماری‌های مختلف با هدف فراهم کردن روش‌هایی دقیق‌تر برای تشخیص، پیشگیری و درمان بیماری‌ها به‌کار می‌رود.

نتایج: استفاده از DNA میکرواری در اندازه‌گیری میزان بیان هزاران ژن به‌طور همزمان، جهت مطالعه رفتار و عملکرد سلول‌ها منجر به تولید حجم انبوهی داده ارزشمند شده است. از این‌رو نقش علم آمار در طراحی، تلخیص، تحلیل و تفسیر درست نتایج این گونه مطالعات اهمیت فوق‌العاده‌ای یافته است. در مقاله حاضر، داده‌های بیان ژنی لوکمیا به‌وسیله روش‌های آماری تجزیه و تحلیل شده است. براساس بررسی انجام شده لوکمیای حاد طبقه‌بندی گردیده، و روش DNA میکرواری، کاربرد آن و برخی روش‌های تحلیل آماری آن‌ها معرفی شده است. در این مطالعه از داده‌های لوکمیا که در سال ۱۹۹۹ توسط Golub منتشر شده استفاده گردیده است.

واژه‌های کلیدی: ریز آرایه‌ها(DNA میکرواری)، بیان ژن، سرطان، طبقه‌بندی سرطان، لوکمیا، روش‌های آماری، Bioconductor

دوماهنامه علمی – پژوهشی دانشگاه شاهد
سال دوازدهم – شماره ۵۴
دی ۱۳۸۳

مقدمه

سرطان‌هایکی از عمده‌ترین نگرانی‌های حال حاضر بشر در حیطه بهداشت عمومی در سراسر جهان محسوب می‌شوند که بعد از حملات قلبی دومین عامل مرگ و میر بشر به‌شمار می‌آیند [۱]. با وجود سیر نزولی روند مرگ و میر مربوط به سرطان به علت تشخیص بیماران در مراحل اولیه، هنوز بسیاری از سرطان‌ها را نمی‌توان به خوبی درمان کرد و اکثر روش‌های موجود برای درمان سرطان نیز توجه خود را معطوف به معالجه بیمارانی می‌کنند که غالباً در مراحل انتهایی قرار دارند.

از این‌رو نیاز به روش‌های جدید به‌منظور تشخیص سرطان قبل از ظهور علائم بالینی، معالجه مؤثرتر بیماران سرطانی که در مراحل اخر بیماری هستند و پیش‌بینی نحوه پاسخ یک تومور به یک روش درمانی پیش از تجویز آن به بیمار کاملاً ضروری به نظر می‌رسد. دانشمندان کلید رسیدن به این اهداف عالیه را منوط به فهم چگونگی عملکرد ژن‌ها و پروتیین‌هایی می‌دانند که در شروع و پیشرفت سرطان‌ها نقش اصلی را به عهده داشته و واکنش بیماران به یک درمان خاص را تحت تأثیر عملکرد خود قرار می‌دهند.

از طرفی جهت درمان مؤثر سرطان‌ها انجام یک طبقه‌بندی دقیق و قابل اعتماد از تومورها ضروری به نظر می‌رسد [۲] امروزه با وجود پیشرفت‌های حاصل در زمینه طبقه‌بندی تومورها که عموماً بر مبنای مشخصات مورفولوژی سلولی و علائم و نشانه‌های بالینی است در این راستا با عدم اطمینان‌های عمده‌ای مواجه می‌باشیم [۳]. با وجود قرار گرفتن افراد بیمار در یک کلاس خاص از طبقه‌بندی، نتایج بالینی متفاوتی در برابر یک روش درمانی واحد مشاهده می‌گردد. همین مسأله، نیاز به انجام طبقه‌بندی‌های دقیق‌تر و جدیدتری از سرطان‌هارا طلب می‌کند. با ظهور تکنولوژی DNA میکرواری امید است با تشخیص تغییرات بیان ژن‌ها و تأثیرات متقابل آن‌ها، درک صحیح‌تری از مکانیزم‌های عمل سلولی و متعاقب آن از فرایندهای ایجاد بیماری‌ها از جمله سرطان حاصل گردد در این مطالعه ضمن تجزیه و تحلیل آماری داده‌های DNA میکرواری مربوط به لوکمیای حاد که توسط گلوب (Golub) در سال ۱۹۹۹ منتشر شده است [۳]، کارایی این تکنیک جهت تشخیص به موقع بیماران و توانایی آن در ارائه طبقه‌بندی معتبر و دقیق‌تری از سرطان‌هانشان داده شده است.

بیان ژنی

هر ژن خود قطعه‌ای از DNA است که بنا بر اصل مرکزی در صورتی بیان می‌شود که ابتدا در هسته قطعه DNA متناظر با آن ژن به mRNA رونویسی شده سپس در سیتوپلاسم این mRNA به پروتیین ترجمه شود.

DNA ⟶ mRNA ⟶ aa(پروتیین)

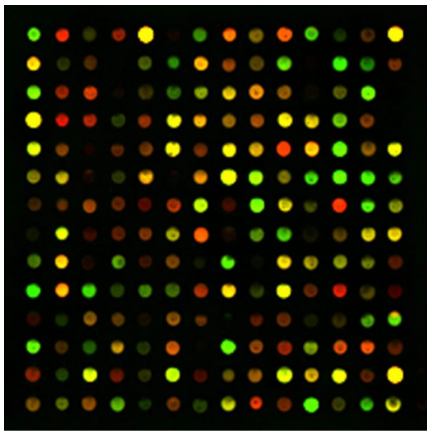
پروتیین‌ها واحدهای عملکردی سلول‌ها هستند که از طریق ترجمه mRNAهای اختصاصی که از روی ژن‌های مربوطه رونویسی شده‌اند ساخته می‌شوند. هر چند DNA موجود در کروموزوم‌های سلول‌های بدن یک فرد یکسان است و به‌عبارتی تمام سلول‌های بدن فرد دارای ژن‌های یکسان هستند. اما سلول‌ها با توجه به نحوه تمایزشان عملکرد متفاوتی نشان می‌دهند به‌عنوان مثال سلول‌های چشم ژن‌های لازم برای دیدن و سلول‌های بینی ژن‌های لازم برای بوییدن را بیان می‌کنند که نحوه بیان متفاوت ژن‌ها در سلول‌های چشم و بینی دلیل عمده این تفاوت عملکرد است.

بیان ژنی و میکرواری

میزان mRNA ی رونویسی شده از یک ژن، با مقدار پروتیین مربوطه تولید شده در سیتوپلاسم سلول رابطه‌ای مستقیم دارد. بنابراین با اندازه‌گیری مقدار mRNA رونویسی شده از یک ژن در سلول می‌توان به‌طور غیرمستقیم میزان بیان آن ژن و میزان پروتیین مربوطه را محاسبه کرد. تکنولوژی DNA میکرواری که در سال ۱۹۹۵ برای اولین بار توسط اسچنا (Schena) [۴] معرفی گردید فرصت گرانبهایی فراهم آورده است تا دانشمندان با اندازه‌گیری میزان mRNA ی تولید شده در هر سلول بتوانند بیان اختصاصی هزاران ژن را بسنجند. از همین‌رو، این تکنولوژی به کانون تحقیقات در فرایندهای سلولی مرتبط با میزان و نحوه بیان ژن از جمله عملکرد ژن‌ها و مکانیزم‌های تمایز سلولی و سرطان تبدیل شده است. به‌وسیله این روش جدید می‌توان با مطالعه همزمان نحوه بیان تمامی ژنوم یک ارگانیسم بر روی یک چیپ، تصویر دقیق‌تری از عملکرد متقابل ژن‌ها را به‌دست آورد.

جهت انجام این تکنیک ابتدا در هزاران لکه (spot) موجود در لام مخصوص (چیپ)، نشانگرهای

اختصاصی هر کدام از ژن‌های مورد مطالعه قرار داده می‌شود. این نشانگرها رشته‌های cDNA یا الیگونوکلئوتیدی مکمل ژن‌های مورد مطالعه هستند. در مرحله بعد mRNAهای نمونه‌های بافتی مورد مطالعه مانند بافت‌های سرطانی و غیره سرطانی بیماران را استخراج کرده و جهت جلوگیری از تجزیه شدن به کمک آنزیم رونویسی معکوس از روی آن‌ها cDNA ساخته و با رنگ‌های فلورسنت نشاندار می‌شوند. مثلاً cDNAهای سرطانی را با قرمز و cDNAهای غیرسرطانی را با سبز نشاندار می‌کنند. در مرحله بعد این cDNAهای نشاندار شده طی فرایند هیبریدیزاسیون به چیپ فوق‌الذکر اضافه می‌شوند و در انتها توسط یک کامپیوتر، چگالی رنگ‌های فلورسنت به‌دست آمده از دستگاه اسکنر برای هر کدام از لکه‌های روی چیپ سنجیده می‌شود و داده‌های به‌دست آمده به‌عنوان میزان بیان ژن تفسیر می‌گردد. با مقایسه نمونه‌های سرطانی و غیرسرطانی هر فرد میزان افزایش یا کاهش بیان هر ژن گزارش می‌شود [۵].



شکل ۱: خروجی نهایی اسکنر که در آن نقاط قرمز نقاطی هستند که ژن‌های سرطانی از بیان بیش‌تری برخوردارند و نقاط سبز بیانگر فراوانی بیش‌تر ژن‌های فرد سالمند و نقاط سیاه نقاطی هستند که در آن‌جا هیچ یک از ژن‌ها بیان نشده‌اند.

داده‌های لوسمی

در این مطالعه از داده‌های بیان ژنی لوکمیای حاد جهت طبقه‌بندی استفاده شده است. لوکمیای حاد به‌طور عمده به دو طبقه لوکمیای حاد (AML) و لوکمیای لیمفوبلاستیک حاد (ALL) تقسیم می‌گردد. این دو طبقه در سطح میکرسکوپی کاملاً یکسان بوده و طی سال‌ها همانند هم عمل می‌کنند [۳]. بنابراین یک تشخیص درست بین این طبقات و بیماران کاملاً ضروری به نظر می‌رسد تا بتوان به بیماران رژیم درمانی مناسب و متفاوتی را تجویز کرد. گرچه در حال حاضر می‌توان به خوبی ALL را از AML تشخیص داد، اما لازمه این تشخیص نیاز به یک متخصص هماتوپاتولوژیست کاملاً خبره و با تجربه جهت تفسیر مورفولوژی تومورها و... و لابراتوارهایی فوق‌العاده تخصصی است. با این همه در حال حاضر چندین ژنی که به‌عنوان مارکرهای لوسمی حاد شناخته می‌شوند کاملاً معلومند لذا این بیانگر آن است که احتمالاً تشخیص بین طبقات لوسمی حاد با استفاده از روش‌های مبتنی بر بیان ژنی قابل پیش‌بینی خواهد بود.

داده‌های لوکمیای استفاده شده در این مطالعه توسط گلوب و همکاران در سال ۱۹۹۹ منتشرگردیده است [۳]. این داده‌ها از ۳۸ نمونه لوکمیا (۱۱ نمونه AML و ۲۷ نمونه ALL) که از مغز استخوان (BM) یا از نمونه خون(PB) افراد گرفته شده به‌دست امده است. و در آدرس اینترنتی http://www.genome.wi.mit.edu/MPR قابل دسترسی است.

مواد و روش‌ها

در غالب مطالعات صورت گرفته بر مبنای اندازه‌گیری‌میزان بیان ژن سرطان‌هابه‌وسیله DNA میکرواری [۳،۷،۸و۹] کشف طبقات (class discovery)، پیش‌بینی طبقات (class prediction) و شناخت مارکرهای ژنی (gene identification) به‌عنوان سه هدف اصلی این تحقیقات مد نظر است. کشف طبقات به شناخت طبقات یا زیر طبقاتی مربوط می‌شود که تاکنون نا

شناخته مانده‌اند، در حالی‌که پیش‌بینی طبقات مربوط به اختصاص دادن بیماران به طبقات موجود و معلوم است. شناخت مارکرهای ژنی به انتخاب زیر مجموعه‌ای از ژن‌ها مربوط می‌گردد که در یک نمونه داده شده از بیان نسبتاً بالاتر یا پایین‌تری برخوردارند.

کشف طبقات

از آن‌جا که رابطه بین ژن‌هادر نمونه‌های آزمایشی دارای ماهیتی چند متغیره است، بنابراین در کشف طبقات روش‌های *کلاسترینگ* به منظور یافتن خوشه‌هایی از نمونه‌ها (یا ژن‌ها) که از الگوی بیان مشابهی برخوردارند استفاده فراوان دارد [۳،۷و۸]. در روش‌های کلاسترینگ نمونه‌ها یا ژن‌هایی که دارای تشابه بیش‌تری (از نظر بیان) با یکدیگر باشند در یک کلاستر یا خوشه قرار می‌گیرند. بدین‌منظور ما در گام نخست قبل از انجام کلاسترینگ مفهوم «تشابه» را از نظر آماری توصیف می‌کنیم. در آمار جهت اندازه‌گیری میزان تشابه چندین معیار وجود دارد که در مطالعات بیان ژن‌ها استفاده از سه معیار ۱) فاصله اقلیدسی، ۲) مربع اختلافات و ۳) ضریب همبستگی پیرسون (r_{ij}) بسیار معمول است. معمولاً ضریب همبستگی جهت بررسی شدت و میزان ارتباط بین دو صفت به‌کار می‌رود. به‌عنوان مثال، در این مورد خاص که هدف ما بررسی ارتباط N بیمار در یک نمونه با یکدیگر براساس بیان ژنی آن‌ها است، هرچه میزان این شاخص (r_{ij}) بیش‌تر باشد حاکی از آن است که دو بیمار مورد نظر یعنی بیماران i و j ($i,j=1,2,\cdots,N$) با یکدیگر وابستگی بیش‌تری دارند. به بیان دیگر چنانچه ضریب همبستگی پیرسون دو بیمار زیاد باشد می‌توان آن دو را متعلق به یک خوشه فرض کرد. ما در این‌جا مقدار کمیت $1-r_{ij}$ را به‌عنوان میزان فاصله دو بیمار براساس بیان تعداد p ژن در نظر گرفته که به‌صورت زیر محاسبه می‌شود به‌کار برده‌ایم:

$$r_{ij}=\frac{\sum_{k=1}^{P}(e_{ik}-\bar{e}_i)(e_{jk}-\bar{e}_j)}{\sqrt{\sum_{k=1}^{P}(e_{ik}-\bar{e}_i)^2\sum_{k=1}^{P}(e_{jk}-\bar{e}_j)^2}}$$

در این‌جا e_{ik} اندازه بیان ژن k ام ($k=1,2,\cdots,P$) در بیمار i ام و $\bar{e}_i=\frac{\sum_{k=1}^{p}e_{ik}}{p}$ میا نگین بیان P ژن بیمار i ام است. همچنین از میان روش‌های کلاسترینگ، از روش خوشه‌بندی سلسله مراتبی جهت تحلیل داده‌های موجود استفاده شده است.

خوشه‌بندی سلسله مراتبی یکی از متداول‌ترین روش‌های کلاسترینگ است که نمونه‌ها (بیماران) را به صورت یک ساختار درختی مرتب می‌کند به‌طوری‌که در این ساختار هر چه نمونه‌ها به هم نزدیک‌تر باشند بیانگر مشابهت بیش‌تر آن‌ها است. این روش بر مبنای معیار فاصله مورد نظر بیماران را در خوشه‌های یکسانی دسته‌بندی می‌کند.

چنانچه $D=(1-r_{ij})=(d_{ij})$ ماتریس فاصله مورد نظر باشد به‌طوری که هر یک از عناصر آن طبق رابطه فوق به‌دست آمده باشند این روش بیماران را به‌صورت زیر گروه‌بندی می‌نماید:

۱- با N خوشه شروع می‌کنیم یعنی در ابتدا هر بیمار به‌عنوان یک خوشه در نظر گرفته می‌شود.

۲- ماتریس فاصله $D_{N\times N}$ را محاسبه می‌نماییم.

۳- d_{ij}، کوچک‌ترین مقدار ماتریس D را پیدا کرده، سپس خوشه جدید $C(i,j)$ از ادغام خوشه $C(i)$ و $C(j)$ ساخته می‌شود.

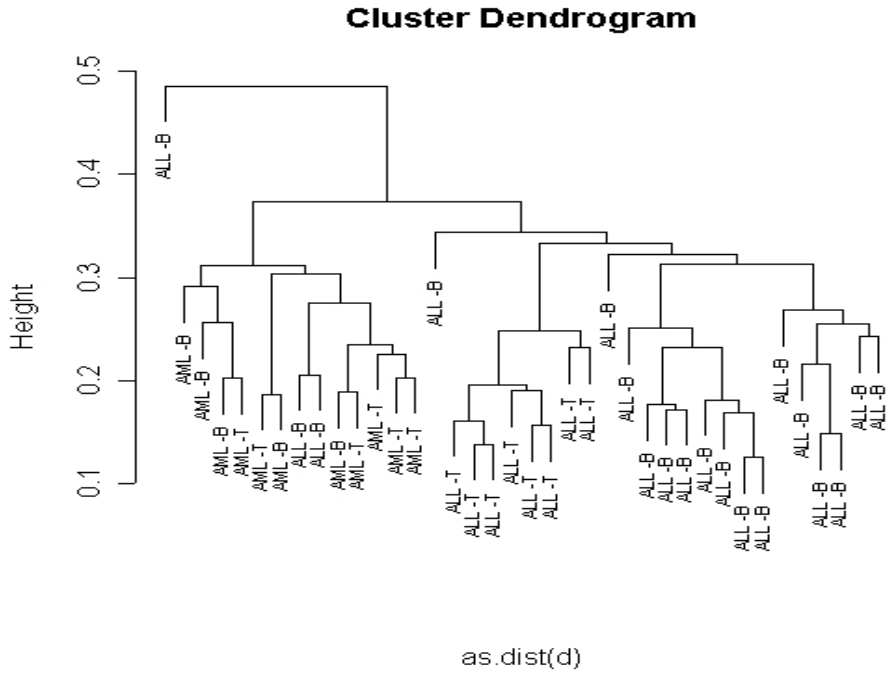
۴- سطر i و ستون j را حذف کرده و ستون جدید $i\cup j$ را می‌سازیم. فاصله $C(i,j)$ و خوشه‌های باقیمانده را با استفاده از یکی از روش‌های single، complete یا average به‌ترتیب زیر به روز می‌شود تا ماتریس فاصله جدید $D_{(N-1)\times(N-1)}$ برای *N-1* عضو جدید به‌دست آید.

$$single = d_{C(i,j),C(k)} = \min\{d_{ik}, d_{jk}\}$$

$$complete = d_{C(i,j),C(k)} = \max\{d_{ik}, d_{jk}\}$$

$$\text{average} = d_{C(i,j),C(k)} = \frac{d_{ik}+d_{jk}}{n_{ij}n_k}$$

۵- گام ۳ و ۴ را *N-1* بار تکرار کرده تا نهایتاً از تمام نمونه‌ها یک خوشه واحد حاصل آید.



نمودار ۱: نمودار درختی خوشه‌بندی بیماران لوسمی با روش سلسله مراتبی

نتایج

در میان روش‌های تجزیه کلاستر به‌منظور تحلیل الگوهای موجود در داده‌های بیان ژن‌ها، روش کلاسترینگ سلسله مراتبی، روشی است که استفاده فراوان دارد. از این روش چندین ویرایش وجود دارد که پرکاربردترین ویرایش آن *ادغام میانگین‌ها* است و ما نیز این ویرایش از روش کلاسترینگ را جهت کشف طبقات موجود در داده‌های لوسمی به‌کار بردیم که حاصل آن نمودار ۱ است.

این نمودار دندوگرام ۳۸ فرد مبتلا به لوسمی حاد را که در آن ۱۱ نمونه لوسمی حاد میلوسیتی با AML و ۲۷ نمونه لیمفوبلاستیکی حاد با ALL نشان داده شده‌اند را نمایش می‌دهد. تابع فاصله مورد استفاده ماتریسی است ۳۸ × ۳۸ که هر یک از عناصر آن از محاسبه یک منهای ضریب همبستگی پیرسون ($1-r_{ij}$) بین افراد بر مبنای ۳۰۵۱ ژن باقیمانده از پیش پردازش اولیه داده‌ها به‌دست آمده‌اند. دندوگرام حاضر با استفاده از نرم‌افزار Bioconductor و در محیط R [۱۰و۱۱] انجام شده است. نمودار را می‌توان به صورت زیر تحلیل کرد:

محور عمودی بیانگر فاصله بین خوشه‌ها است که با استفاده از روش ادغام میانگین‌ها به‌دست آمده‌اند و ارتفاع هر یک از این شاخه‌ها بیانگر آن است که دو خوشه در چه نقطه‌ای با هم ادغام شده‌اند. در این محور هر چه از طرف پایین به سمت بالا پیش می‌رویم بیانگر فاصله بیش‌تر خوشه‌ها ست. نتیجه انجام این روش کلاسترینگ در جدول ۱ آورده شده است.

همان‌طوری که در این جدول می بینید نمونه‌ها را می‌توان در سه طبقه یا خوشه دسته‌بندی کرد. طبقه اول از ۱۶ بیمار ALL-B و ۸ بیمار ALL-T، طبقه دوم از ۲ بیمارALL-B و ۱۱ بیمار AML و بالاخره خوشه سوم که تنها شامل ۱ نمونه از فرد مبتلا به ALL می باشد. ابتدا ما به ارزیابی خوشه‌های حاصل به‌وسیله تعلق داشتن آن‌ها به طبقات معلوم AML-ALL می‌پردازیم. روش سلسله مراتبی در تخصیص افراد به دو طبقه معلوم ALL و AML کاملاً تواناست زیرا خوشه اول در برگیرنده ۲۴ نمونه ALL از ۲۵ نمونه موجود و طبقه دوم دربرگیرنده تمام ۱۱ نمونه AML است که به جز در

دو مورد تخصیص‌ها کاملاً درستند. نتایج این روش طبقه‌بندی در جدول ۲ آورده شده است.

جدول ۱: نتایج انتساب بیماران لوسمی در خوشه‌هابا استفاده از روش سلسله مراتبی

نوع لوسمی	شماره خوشه ها			جمع
	۱	۲	۳	
ALL-B	۱۶	۲	۱	۱۹
ALL-T	۸	۰	۰	۸
AML	۰	۱۱	۰	۱۱
جمع	۲۴	۱۳	۱	۳۸

جدول ۲: نمایش حساسیت روش خوشه‌بندی در تعیین نوع لوسمی

انتساب پس از خوشه‌بندی	نوع لوسمی		جمع
	ALL	AML	
ALL	۲۵	۰	۲۵
AML	۲	۱۱	۱۳
جمع	۲۷	۱۱	۳۸

از این جدول می‌توان نتیجه گرفت که ویژگی [(شخص سالم | p(test -] روش سلسله مراتبی در تشخیص افراد ALL برابر ۹۳ درصد (۲۵ نفر از ۲۷ نفر) و حساسیت [(شخص بیمار | p(test +] آن برابر ۱۰۰ درصد (۱۱نفر از ۱۱ نفر) است. بنابراین دقت کلی این‌روش برابر با ۹۵ درصد خواهد بود. به عبارت دیگر این روش چنانچه فردی را ALL تشخیص دهد این فرد یقیناً مبتلا به سرطان لوکمیای میلوسیتی حاد خواهد بود. بنابراین این‌روش در کشف اتوماتیک لوسمی به دو نوع حاد آن کاملاً (هر چند به‌طور ناتمام) مؤثر است.

حال اگر بخواهیم در نمودار ۱ یک مرحله پایین‌تر آمده و جستجوی خود را برای کشف طبقات جدیدتر و بهتری از لوسمی ادامه دهیم، می‌توان از خوشه اول که شامل ۲۴ نمونه ALL است دو زیر طبقه دیگر، شامل ۸ نمونه ALL-T و ۱۶ نمونه ALL-B را تمییز داد. این نتیجه می‌تواند بسیار حائز اهمیت باشد زیرا در هنگام شروع تحقیق ما این نمونه‌ها را فقط به‌عنوان افرادی می‌شناختیم که تنها دچار یکی از دو نوع سرطان لوکمیای ALL یا AML بودند. تفکیک این دو زیرطبقه از ALL نتیجه دقیق‌تری است که انجام این‌روش کلاسترینگ بر روی داده‌های بیان ژنی حاصل از DNA میکروواری منجر به کشف آن شده است. بنابراین روش کشف طبقات نه تنها به‌طور اتوماتیک AML را از ALL تشخیص داد بلکه قادر به تمایز بین ALL-T و ALL-B نیز است. این احتمال وجود دارد که با نمونه‌های بیش‌تر هنوز بتوان زیر طبقات بهتری را شناخت.

بحث

نتایج به‌دست آمده از DNA میکروواری منجر به ایجاد دیدگاه‌های تازه‌ای در مورد نحوه شکل‌گیری، پیشرفت و پاسخ به درمان بیماران سرطانی گردیده است. با توجه با این‌که سکانس کامل ژنوم انسانی در دسترس است بررسی کامل نسخه‌برداری در سلول‌های نرمال و سرطانی امکان‌پذیر گردیده است و همراه با تکامل همزمان ابزارهای ضروری انفورماتیک و آنالیز داده‌ها جهت تبدیل و تفسیر آن‌ها، نحوه نگرش به سرطان دچار تحول شدیدی گردیده است. ترکیبی از روش‌های ژنومیک و پروتئومیکس احتمالا سبب پیشرفت‌های عمیق‌تری در این زمینه خواهد شد. تجزیه کلاستر برای مسائلی طرح‌ریزی شده است که با در دست داشتن نمونه‌ای از n فرد و اندازه‌گیری p صفت (در این‌جا ژن‌ها) برای هر فرد، بتوان افراد را در طبقاتی گروه‌بندی کرد که افراد دارای بیش‌ترین تشابه در داخل یک طبقه قرار گیرند. در این روش داده‌هابه صورت یک نمودار درختی نمایش داده می‌شوند و الگوهای مشابه در یک زیرطبقه دسته‌بندی می‌شوند. دلایل زیادی را می‌توان برای نشان دادن ارزشمند بودن تجزیه کلاستر ارائه داد. اولاً تجزیه کلاستر می‌تواند در یافتن گروه‌های واقعی و نهفته در داده‌ها کارساز باشد. ثانیاً تجزیه کلاستر می‌تواند در کاهش داده‌هامؤثر باشد. به‌عنوان مثال، به‌منظور تأثیر یک روش درمانی باید تعداد زیادی از افراد را مورد آزمایش قرار داد اما با

توجه به محدودیت‌های موجود تنها تعداد کمی از افراد مورد آزمایش قرار می‌گیرند. حال اگر بتوان افراد را به تعداد کم‌تری از طبقات مشابه گروه‌بندی کرد در این صورت از هر طبقه یک فرد می‌تواند به‌عنوان نماینده آن طبقه انتخاب شود. ثالثاً تجزیه کلاستر ممکن است منجر به کشف طبقات جدید و غیرقابل انتظاری شود که در این صورت نتیجه حاصل بیانگر روابط جدیدی خواهد بود.

روش‌هایی که در این‌جا مورد بحث قرار گرفت، مانند کلاسترینگ و طبقه‌بندی فرض‌های کلی بودند. فرضیات ویژه و نهایی در مطالعات مختلف می‌توانند بسیار متنوع باشند به همین علت در انتخاب روش‌های آماری مناسب جهت مطالعه افراد باید دقت زیادی را صرف کرد. در واقع روش فوق جهت کشف طبقات را می‌توان جهت مشخص کردن زیر طبقات احتمالی در مورد هر نوع سرطان دیگری را نیز به کار برد. روش کشف طبقات را می‌توان همچنین جهت تحقیق پیرامون مکانیزم‌های اساسی که باعث تشخیص انواع سرطانها را به‌کار بست. به‌عنوان مثال، می‌توان سرطان‌های متفاوتی را با هم در یک مجموعه واحد ترکیب کرد (مانند، تومور پستان و نومور پروستات) و بعد از حذف ژن‌هایی که با نوع بافت‌های مورد نظرهمبستگی بالایی دارند، نمونه‌هارا براساس ژن‌های باقیمانده طبقه‌بندی کرد.

ما یک روش پیش‌بینی‌کننده طبقات را معرفی کردیم که نه تنها قادر به تشخیص درست AML از ALL است بلکه به درستی قادر به تخصیص ALL به ALL-B و ALL-T بود. بنابراین می‌توان این‌روش واحد را به جای چندین روش موجود که هم به صرف هزینه و هم نیاز به متخصصانی خبره و با تجربه دارند جایگزین کرد. در حقیقت این تکنولوژی 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. از طرفی روش پیش‌بینی طبقات را می‌توان جهت تشخیص نتایج بالینی در آینده، از قبیل واکنش به یک دارو یا میزان بقای افراد بیمار را نیز به کار برد.

لازم به ذکر است که نتایجی که ما در این‌جا با استفاده از روش کلاسترینگ سلسله مراتبی به‌دست آوردیم با نتایج گلوب و همکارانش [۳] که با استفاده از تکنیک خوشه‌بندی دیگری موسوم به SOM به‌دست آورده‌اند همخوانی کامل دارد.

از روش‌های بسیار متداول آماری در تحلیل داده‌های بیان ژنی روش‌هایی هستند که در تشخیص ژن‌های مسئول بیماری‌ها استفاده می‌گردد. همان‌گونه که معلوم است تشخیص چندین ژن مسول از میان انبوه ژن‌ها با استفاده از روش‌های کلاسیک آماری و مقایسه p-value های محاسبه شده امکان‌پذیر نبوده و نیاز به تعدیلp-valueها است. در این راستا Efron، Tibshirani و دیگران راهکاره‌هایی را بر مبنای آمار بیزی و بیزی تجربی ارائه کرده‌اند که به روش‌های FDR معروفند [۱۲و۱۳]. در حالت کلی این روش‌ها برای انجام مقایسات سطح معناداری خطای نوع اول α را در تمام مقایسات ثابت نگه‌می‌دارند و بر این‌اساس p-value های تعدیل شده‌ای را تولید می‌کنند. علاوه بر روش‌های FDR روش‌های دیگری نیز وجود دارند که بدون بهره‌گیری از آمار بیز براساس الگوریتمی ساده‌تر p-value های تعدیل شده‌ای را ارائه کرده‌اند. این‌روش‌ها که به روش‌های FWER [۱۴و۱۵] موسومند به‌صورت زیر عمل می‌کنند:

۱- داده‌هالوسمی را در قالب یک ماتریس ۳۸ ستونی ساخته سپس این ستون‌ها را به صورت تصادفی یا سیستماتیک عوض می‌کنیم تا در هر مرحله یک جایگشت به‌دست آید.

۲- چنانچه بخواهیم برای مقایسه‌های دو به دو از آماره t معمولی استفاده کنیم، برای هر یک از ژن‌ها آماره‌های- t را در هر جایگشت (b) محاسبه می‌کنیم:

$t_1^{(b)},\ldots,t_k^{(b)}$

۳- قرار می‌دهیم:

$u_k^{(b)} \geq \left|t_{rj}\right|$

$$u_j^{(b)} = \max\left(u_{j+1}^{(b)}, \left|t_{rj}^{(b)}\right|\right) \qquad 1 \le j \le N-1$$

در این جا r_j به نحوی هستند که در داده‌های اولیه $|t_{r1}| \ge |t_{r2}| \ge \cdots \ge |t_{rk}|$ برقرار باشد.

۴- گام ۲ تا ۳ را B بار تکرار می‌کنیم تا P- Value‌های تعدیل شده به‌صورت زیر به‌دست آیند.

$$\widetilde{P}_{rj}^{\,*} = \frac{\sum_{b=1}^{B} I\left(u_j^{(b)} \ge \left|t_{rj}\right|\right)}{B}$$

۵- به شرط یکنوایی P- Value‌ها

$$\widetilde{P}_{r1}^{\,*} \leftarrow \widetilde{P}_{r1}^{\,*}$$

$$\widetilde{P}_{rj}^{\,*} \leftarrow \max\left(\widetilde{P}_{rj}^{\,*}, \widetilde{P}^{*}_{rj-1}\right) \qquad 2 \le j \le k$$

در نهایت ژن‌هایی را که P- Value تعدیل شده‌ای کم‌تر از 05. داشته باشند به‌عنوان ژن‌هایی که بیان متفاوتی دارند (ژن های مسئول بیماری) گزارش می‌شوند.

امروزه تکنیک DNA میکرواری به‌عنوان یک روش جایگزین برای تشخیص بیماری‌ها و در ساختن دارو مطرح‌است. در بسیاری‌از بیمارستان‌ها و مراکز تحقیقاتی امریکا و اروپا از دستگاه Affymetrix به‌صورت تحقیقاتی یا تجاری با استفاده از داده‌های روزمره بیمارستانی بیماران استفاده می شود. همان‌گونه که اطلاعات انبوه DNA میکرواری افق‌های‌روشنی را جهت درمان بیماری‌ها ترسیم می‌کند به علت پیچیدگی خاص اطلاعات انبوه حاصل از این دستگاه‌ها، تحلیل‌های آماری در مورد این داده‌ها متفاوت و پیچیده‌تر است. روش‌های که تاکنون به‌وسیله آماردانان برای تحلیل این داده‌ها پیشنهاد شده هنوز در مرحله اولیه بوده و در حال توسعه روزافزون است [۱۵]. برای پیشبرد اهداف تحقیقانی در این مقوله همکاری بین پزشکان، متخصصان بیوشیمی و بیولوژی مولکولی و متخصصین آمار بیش از همیشه مورد توجه قرار گرفته است.

منابع

1. Ochs MF, Godwin AK. Microarray in cancer: Research and applications, Bio Techniques, 2003; 34:4-15.
2. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data, JASA, Technical reports, http://www.stat.berkeley.edu/users/sandrine, 2002.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et. al. Molecular classification of cancer:class discovery and classification by gene expression monitoring. Science, 1999; 286:531-537.
4. Schena M, Shalon D, Davis RW, Brown PO, Quantitative monitoring of gene expression patterns with a complementary DNA microarray cancer. Science, 1995; 270:467-476.
5. Nguyen V, Danh AB, Arpat NW, Carroll RJ, DNA Microarray Experiments: Biological and Technological Aspects, Biometrics, 2002.
6. Slonim DK, Tamayo P, Messirov JP, Golub TR, Lander ES, Class prediction and discovery using gene expression data., ACM, 2000.
7. Alizadeh AA, et. al., Different types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature, 2000; 403:503-511.
8. Alon, U, Barkai N, Notterman DA, et. al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci., 1999; 96:6745-6750.
9. Eisen M, Spellman PL, Brown PO, Botstain D, Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci., 1998; 95:14863-14868.
10. http://www.bioconductor.org.
11. http://www.r-project.org.
12. Efron B, Tibshirani R, Storey D, Virginia Tusher V, Empirical bayes analysis of a microarray experiment , Journal of American Statistical Association, 2001.
13. Speed T, Statistical analysis of gene expression microarray data, Chapman & Hall/CRC, 2003.
14. Dudoit S, et al. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of American Statistical Association. Vol.97. Technical Report # 576, Department of Statistics, U. C. Berkeley, 2002.
15. آزادی نمامعلی، روش‌های آماری در تحلیل داده‌های بیان ژنی به‌دست آمده از تکنولوژی DNA میکرواری، پایان‌نامه کارشناسی ارشد، دانشکده بهداشت، دانشگاه علوم پزشکی تهران.