پژوهنده (مجله پژوهشی دانشگاه علوم پزشکی شهید بهشتی)
سال چهاردهم، شماره ۶، پی در پی۷۲، صفحات ۲۸۸ تا ۲۹۴
بهمن و اسفند ۱۳۸۸

تاریخ دریافت مقاله: ۱۳۸۷/۱۱/۹
تاریخ پذیرش مقاله: ۱۳۸۸/۹/۱۱

به‌کارگیری خوشه‌بندی فازی در ریزآرایه DNA

محسن واحدی [۱*]، دکتر حمید علوی‌مجد [۲]، دکتر یدا... ممرابی [۳]، بهار نقوی [۴]

۱. کارشناس ارشد آمار زیستی، مرکز تحقیقات بیماریهای گوارش و کبد، دانشگاه علوم پزشکی شهید بهشتی
۲. دانشیار، گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی
۳. استاد، گروه آمار و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی شهید بهشتی
۴. مربی، دانشکده پرستاری و مامایی، دانشگاه علوم پزشکی شهید بهشتی

چکیده

سابقه و هدف: فناوری ریزآرایه برای بررسی همزمان بیان هزاران ژن در بازه وسیعی از ژنومیک، نظیر شناسایی ژنها، اکتشاف داروها و تشخیص‌های کلینیکی مورد استفاده قرار گرفته است. آزمایشهایی که بر اساس فناوری ریزآرایه انجام می‌شوند حجم بسیار زیادی از داده‌ها را فراهم می‌کنند که در مطالعات بیولوژیک بی‌نظیر بوده است. نرمال سازی، خوشه‌بندی، طبقه‌بندی و ... از جمله روشهای مورد استفاده در تحلیل آماری این نوع داده‌هاست. هدف این مقاله بررسی نحوه به کارگیری روش خوشه‌بندی فازی در داده‌های ریزآرایه DNA است.

مواد و روش‌ها: تحقیق به روش توصیفی انجام شده و 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 روش آماری خوشه‌بندی فازی، مورد تجزیه و تحلیل قرار گرفته است. خوشه‌بندی داده‌های ریزآرایه به صورت خیلی محدود و گرایش بیشتر به سمت خوشه‌بندی کلاسیک در ایران صورت پذیرفته است و این مقاله شروعی در خوشه‌بندی فازی داده‌های ریزآرایه به شمار می‌آید. مجموعه داده‌ها شامل ۲۰ بیمار مبتلا به سرطان خون لنفوئیدی حاد (ALL) و ۱۴ بیمار مبتلا به سرطان خون میلوئیدی حاد(AML) است. کارایی روش خوشه‌بندی فازی با توجه به گروه‌بندی واقعی نمونه‌ها (ALL و AML) مورد ارزیابی قرار گرفت. نرم‌افزار R برای تحلیل داده‌ها استفاده شد.

یافته‌ها: ویژگی روش خوشه‌بندی فازی در تشخیص افراد AML، ۹۰٪ و حساسیت آن ۹۳٪ و صحت آن ۹۱٪ به دست آمد که نشان‌دهنده عملکرد خوب این روش است. نمونه سی و یکم که بر اساس یافته‌های بالینی در گروه AML قرار دارد در گروه ALL قرار گرفت، همچنین نمونه‌های دوم و هفدهم که بر اساس یافته‌های بالینی در گروه ALL قرار دارد طبق نتایج در گروه AML قرار گرفتند که از نظر بالینی می‌توانند قابل توجه باشند.

نتیجه‌گیری: خوشه‌بندی فازی اطلاعات نسبتاً قابل قبولی درباره ساختار داده‌ها فراهم می‌کند که با توجه به انطباق نتایج این روش با گروه‌بندی واقعی داده‌ها، از این روش آماری می‌توان در مواردی که اطلاع دقیقی درباره گروه‌بندی واقعی داده‌ها در دست نیست، استفاده کرد. به علاوه با بررسی نتایج خوشه‌بندی ممکن است زیرگروه‌هایی از نمونه‌ها را به نحوی متمایز کرد که برای انطباق آن با یافته‌های بالینی، پژوهشهای آزمایشگاهی یا بالینی جدیدی لازم باشد.

واژگان کلیدی: ریزآرایه DNA، بیان ژن، خوشه‌بندی کلاسیک، خوشه‌بندی فازی، سرطان خون

مقدمه

در سالهای اخیر فناوری ریزآرایه، امکان کنترل بیان هزاران ژن را به صورت همزمان فراهم کرده و ژنومیکس و پروتئومیکس اساس شیوه‌های علمی مطالعه پایه مولکولی رفتارهای بافت و سلول در شرایط فیزیولوژیک و پاتولوژیک را تغییر داده‌اند. فناوری ریزآرایه برای بررسی همزمان بیان هزاران ژن در بازه وسیعی از ژنومیک، نظیر شناسایی ژنها، اکتشاف داروها و تشخیص‌های کلینیکی به صورت موفقیت‌آمیزی مورد استفاده قرار گرفته است (۱).

داده‌های سطوح بیان ژنها اطلاعات ارزشمندی در مورد شبکه‌های بیولوژیک، حالات سلولی و فهمیدن کارکرد ژنها در

* **نویسنده مسئول مکاتبات:** محسن واحدی؛ تهران، اوین، خیابان تابناک، بیمارستان آیت ا... طالقانی، مرکز تحقیقات بیماریهای گوارش و کبد؛ **پست الکترونیک:** mohsenvahedi540@gmail.com

بر دارد. یک هدف از تحلیل داده‌های بیان ژن، تعیین چگونگی تأثیر بیان هر ژن منفرد روی بیان ژنهای دیگر در همان شبکه ژنتیکی است. هدف دیگر، مشخص کردن این نکته است که چگونه ژنها در سلولهای سالم و بیمار بیان می‌شوند. کاربرد علمی بررسی بیان ژن ریزآرایه مدیریت و کنترل سرطان و بیماریهای عفونی است. هدف اصلی این مطالعات، تعیین و شناسایی فرایند پاتولوژیک مرتبط با نوع بیماری و مرحله آن و نیز پیش‌بینی پاسخ به درمان خاصی است. همچنین برخی از مسایل در زمینه تشخیص، با استفاده از تحلیل داده‌های بیان ژن قابل حل شده‌اند (۲).

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 (۳).

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

آنالیز پایه داده‌های بیان ژن شامل یک مرحله پیش پردازش و آماده‌سازی مجموعه داده برای مراحل آنالیز سطوح بالاتر است. پیش پردازش داده‌های خام می‌تواند اثرات عمیقی روی مراحل بعدی آنالیز داشته باشد. استفاده از روشهای آماری برای تحلیل داده‌های حاصل از این فناوری می‌تواند گام مؤثری در جهت تشخیص و درمان بیماریها داشته باشد. یکی از روشهای آماری که در تحلیل این داده‌ها به صورت فزاینده مورد استفاده قرار می‌گیرد خوشه‌بندی است. از سال ۱۹۹۸ خوشه‌بندی داده‌های بیان ژنی شروع گردیده است (۴). در ابتدا فقط روش های خوشه‌بندی کلاسیک نظیر خوشه‌بندی سلسله مراتبی مورد استفاده قرار می‌گرفت که در بسیاری از موارد روشهای کارا و مفید بودند ولی در بعضی از مواقع کاستی‌هایی نیز داشتند که باعث شد تا روشهای خوشه‌بندی غیرکلاسیک نظیر خوشه‌بندی فازی در نظر گرفته شود؛ مثلاً به منظور از بین بردن مشکل خوشه‌بندی اشتباه داده‌های ریزآرایه از روش خوشه‌بندی فازی استفاده شد (۵). متدولوژی استفاده از این روش نیز ارایه گردید (۶). در مواردی که ساختار داده‌ها پیچیده است و خوشه‌بندی‌های رایج جواب نمی‌دهند، روش خوشه‌بندی فازی FCM (Fuzzy C-Means) می‌تواند گره‌گشا باشد (۷). همچنین با مقایسه اعتمادپذیری تحلیل‌های ریزآرایه با توجه به دو روش FCM و مدلهای آمیخته نرمال (Normal Mixture Modeling) در حالت مدلهای آمیخته، با توجه به سرعت بیشتر FCM در چنین تحلیل‌هایی این روش برتر از NMM می‌باشد (۸).

خوشه‌بندی فازی را می‌توان به عنوان حالت تعمیم یافته افراز قطعی (Hard Partitioning) در نظر گرفت. در یک افراز قطعی (خوشه‌بندی سلسله مراتبی متعلق به این رده می‌باشد) هر عنصر فقط به یک خوشه تعلق دارد.

در روش خوشه‌بندی فازی برای هر عنصر میزان قرار گرفتن در خوشه‌های مختلف بوسیله ضرایب عضویت (Membership Coefficients) که اعدادی در بازه ۰ تا ۱ است سنجیده می‌شود. ضریب عضویت عنصر در هر خوشه‌ای که بیشتر بود عنصر متعلق به آن خوشه است.

اصولاً پایه‌ریزی الگوریتم خوشه‌بندی فازی برای فائق آمدن بر وضعیت بروز مشکل همپوشی خوشه‌ها که از طریق وجود عنصرهایی که در تعلق آنها به خوشه‌های مختلف ابهام وجود دارد ابداع شد. اساس آن مبتنی بر تقسیم n آزمودنی به c خوشه از طریق محاسبه و تعیین ضریب عضویت هر آزمودنی به هر خوشه می‌باشد.

امتیاز اصلی خوشه‌بندی فازی نسبت به خوشه‌بندی قطعی ارایه اطلاعات دقیقتری درباره ساختار داده‌ها می‌باشد. از طرف دیگر این ممکن است به عنوان یک اشکال در نظر گرفته شود، زیرا مقدار خروجی با افزایش تعداد عناصر و تعداد خوشه‌ها به سرعت زیاد می‌شود که در نتیجه زمان زیادی برای تحلیل صرف می‌شود. 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 (۹).

هدف این مقاله بررسی نحوه به کارگیری روش خوشه‌بندی فازی در داده‌های ریزآرایه DNA است.

مواد و روش‌ها

این مطالعه از نوع مشاهده‌ای- مقطعی می‌باشد.

فناوری ریزآرایه DNA و روش‌های تحلیل داده‌های آن جدید می‌باشند، بنابراین دراین و خیلی از کشورهای دیگر تحقیقاتی در زمینه تولید این گونه داده‌ها صورت نگرفته است، لذا اغلب محققان ناچارند که از داده‌های بانک‌های اطلاعاتی اینترنتی نظیر GenBank استفاده کنند (۱۰).

در این تحقیق از داده‌های سرطان خون که توسط گلوب و همکاران(Golub) در سال ۱۹۹۹ انتشار یافته استفاده شد (۱۱). 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 لنفوییدی(ALL) و ۱۴ نمونه از بیماران با سرطان خون حاد میلوییدی(AML) می‌باشند، که به صورت نمونه‌گیری مبتنی بر هدف و غیرتصادفی انتخاب شده‌اند و بر اساس روش آرایه الیگونوکلئوتید بیان ژنها به دست آمده است. این داده‌ها از طریق اینترنت در اختیار عموم قرار دارد (۱۲).

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

با توجه به توصیه‌های گلوب و همکاران (۱۹۹۹) موارد ذیل برای پیش‌پردازش 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 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 طول ۳۴ نمونه می‌باشد.

۳) انجام تبدیل لگاریتمی بر روی داده‌ها

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

با انجام این پیش‌پردازش تنها ۲۹۱۷ ژن در ماتریس داده‌ها باقی ماندند و ماتریس بیان ژنی ۳۴×۲۹۱۷ برای تحلیل آماری مورد استفاده قرار گرفت.

تکنیک خوشه‌بندی فازی را با مینیمم کردن تابع هدف زیر به دست آورده‌ایم:

$$C = \sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 \, d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2} \quad \#$$

$d(i,j)$ نشان‌دهنده تفاوت یا فاصله بین دو عنصر i و j است، u_{iv} ضریب عضویت نامعلوم عنصر i به خوشه v است. توابع عضویت به صورت زیر مقید شده‌اند:

$$u_{iv}^2 \ge 0 \quad for \quad i = 1,\ldots,n \; ; v = 1,\ldots,k$$

$$\sum_{v} u_{iv} = 1 \quad for \quad i = 1,\ldots,n$$

که نشان‌دهنده این نکته است که عضویت‌ها نمی‌تواند منفی باشد و مجموع ضرایب عضویت یک عنصر روی خوشه‌ها برابر با یک است. از تابع لاگرانژ که برای بهینه کردن (#) استفاده کردیم.

به خاطر محدودیت‌هایی که برای موردهای فازی است راه حل‌های خوشه‌بندی قطعی مورد نظر قرار می‌گیرد. میزان تفاوت یک راه حل فازی از خوشه‌بندی قطعی را می‌توان از ضریب افراز دان (Dunn's Partition Coefficient) ارزیابی کرد که به صورت مجموع مربعات همه ضرایب عضویت تقسیم بر تعداد عناصر به دست می‌آید.

$$F_k(U) = \sum_{i=1}^{n}\sum_{v=1}^{k} u_{iv}^2 / n$$

که U ماتریس ضرایب عضویت است.

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1k} \\ u_{21} & u_{22} & \cdots & u_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nk} \end{bmatrix}$$

برای یک افراز(u_{iv} محدود شده به صفر ویک) $F_k(U)$ ماکزیمم مقدار یک را می‌گیرد در حالی که مینیمم مقدار $1/k$ را وقتی همه $u_{iv} = 1/k$ است به خود اختصاص می‌دهد. این ضریب می‌تواند به صورتی که از ۱ (خوشه‌های قطعی) تا ۰ (تماماً فازی) تغییر کند، نرمال شود. این کار مستقل از تعداد خوشه‌ها به وسیله تبدیل انجام شود:

$$F_k'(U) = \frac{F_k(U) - (1/k)}{1 - (1/k)} = \frac{kF_k(U) - 1}{k - 1}$$

این ضریب نرمال شده اغلب به عنوان شاخص غیرفازی (Nonfuzziness Index) نامیده می‌شود (۱۳).

اگر بخواهیم یک خوشه‌بندی فازی را به خوشه‌بندی قطعی تبدیل کنیم باید $w_{iq} = 1$ برای خوشه q که بزرگترین مقدارهای u_{iv} را دارد تعریف کنیم. در مواردی که گره وجود دارد به صورت انتخابی عمل می‌کنیم. این تبدیل برای مقایسه فازی با یک راه حل قطعی مورد استفاده قرار می‌گیرد.

برای مقایسه نتایج دو روش فازی و غیرفازی و بررسی این موضوع که نتایج آنها تا چه حد با هم تطابق دارند، لازم است نتایج حاصل از روش غیرفازی با مقادیر و ضرایب عضویت عناصر به خوشه‌های متناظر حاصل از روش فازی مقایسه شود که بالا بودن مقدار ضریب عضویت برای یک عنصر مثل i به خوشه‌ای که مثل v که همین عنصر بر طبق روش غیرفازی در آن قرار گرفته است بیانگر مطابقت نتایج دو روش است.

اجرای روش خوشه‌بندی فازی با استفاده از نرم افزار R دارای عبارات و مفاهیمی است. یکی از این مفاهیم نمودار سایه‌نما (Silhouette) است. نمودار سایه‌نما اولین بار توسط رسیو معرفی گردید (۱۴). در این نمودار هر خوشه به صورت یک سایه‌نما مطرح می‌شود که در آن وضعیت هر آزمودنی که در آن خوشه قرار گرفته است، به لحاظ شدت تعلق به آن خوشه مشخص می‌باشد. کل خوشه‌بندی با قرار گرفتن چند سایه‌نما (به تعداد خوشه‌های تعیین شده) در کنار هم معین می‌گردد و به کاربر این اجازه را می‌دهد که به مقایسه کیفیت خوشه‌های تشکیل شده بپردازد. چنین نمودار سایه‌نمایی در هنگامی که مقیاس داده‌ها نسبی است، بسیار مفید است، با استفاده از آن به راحتی فشرده بودن داخلی هر خوشه و تفکیک خوشه‌های مختلف می‌تواند مورد بررسی قرار گیرد.

برای هر خوشه میانگین مقادیر سایه‌نما را به دست می‌آوریم. بزرگترین مقدار میانگین در بین خوشه‌های مختلف به عنوان ضریب سایه‌نما نامیده می‌شودکه به صورت جدول زیر تفسیر می‌شود (۱۳).

مقدار عددی ضریب سایه‌نما	تفسیر مربوطه
۱-۰/۷۱	ساختار خوشه‌ای بسیار مناسبی تشکیل شده است
۰/۷۰-۰/۵۱	ساختار خوشه‌ای معقولی تشکیل شده است.
۰/۵۰-۰/۲۶	ساختار خوشه‌ای ضعیف است و توصیه می‌شود از الگوریتم دیگری استفاده شود.
۰/۲۵-۰	اصولاً یافتن ساختار مناسب برای این داده‌ها مشکل می‌باشد.

در این تحقیق داده‌های بیان ژنی پس از پیش‌پردازش اولیه با روش خوشه‌بندی فازی، خوشه‌بندی گردید و برای ارزیابی خوشه‌بندی از مقادیر و نمودار سایه‌نما استفاده شد.

یافته‌ها

روش خوشه‌بندی فازی در مورد وضعیت‌های مبهمی که در داده‌ها رخ می‌دهد یعنی حالتی که خوشه‌ها از طریق وجود عنصرهایی که از لحاظ تعلق به خوشه‌های مختلف در یک وضعیت بینابینی قرار می‌گیرند و حالت هم‌پوشی در آنها ایجاد می‌شود، کاربرد دارد. خوشه‌بندی فازی را ابتدا با ۲ خوشه شروع کردیم، سپس ضریب عضویت هر یک از نمونه‌ها به هر خوشه محاسبه گردید. بر اساس این که ضرایب عضویت در کدام خوشه بیشتر باشد می‌توان نمونه را به آن خوشه اختصاص داد، نتایج این اختصاص در جدول ۱ آمده است.

برای بیان میزان تفاوت یک راه حل فازی از خوشه‌بندی قطعی از ضریب افراز دان استفاده کردیم که مقدار آن ۰/۸۷۴ و شاخص غیرفازی ۰/۷۴۸ به دست آمد.

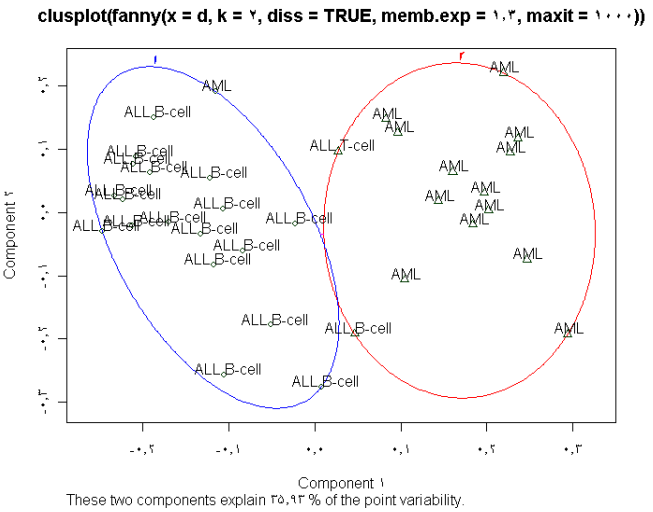
جدول۱- انجام خوشه‌بندی فازی به دو خوشه و مقادیر سایه نما

نمونه	نوع لوسمی	ضرایب عضویت		نزدیکترین خوشه	مقادیر سایه نما
		خوشه ۱	خوشه ۲		
۱	ALL B-cell	۰/۷۲۷	۰/۲۷۳	۱	۰/۴۶۰
۲	ALL B-cell	۰/۴۵۱	۰/۵۴۹	۲	۰/۰۴۱
۳	ALL B-cell	۰/۷۸۸	۰/۲۱۲	۱	۰/۴۵۸
۴	ALL B-cell	۰/۷۸۲	۰/۲۱۸	۱	۰/۴۳۲
۵	ALL B-cell	۰/۹۰۱	۰/۰۹۹	۱	۰/۴۱۴
۶	ALL B-cell	۰/۵۰۹	۰/۴۹۱	۱	۰/۳۹۵
۷	ALL B-cell	۰/۸۸۲	۰/۱۱۸	۱	۰/۳۸۴
۸	ALL B-cell	۰/۷۸۵	۰/۲۱۵	۱	۰/۳۴۵
۹	ALL B-cell	۰/۸۹۹	۰/۱۰۱	۱	۰/۳۳۹
۱۰	ALL B-cell	۰/۸۷۵	۰/۱۲۵	۱	۰/۳۲۴
۱۱	ALL B-cell	۰/۹۱۹	۰/۰۸۱	۱	۰/۳۱۰
۱۲	ALL B-cell	۰/۸۱۶	۰/۱۸۴	۱	۰/۳۰۹
۱۳	ALL B-cell	۰/۷۳۸	۰/۲۶۲	۱	۰/۲۹۱
۱۴	ALL B-cell	۰/۸۰۸	۰/۱۹۲	۱	۰/۲۷۱
۱۵	ALL B-cell	۰/۸۲۵	۰/۱۷۵	۱	۰/۲۶۷
۱۶	ALL B-cell	۰/۸۹۷	۰/۱۰۳	۱	۰/۲۵۳
۱۷	ALL T-cell	۰/۴۳۵	۰/۵۶۵	۲	۰/۰۲۱
۱۸	ALL B-cell	۰/۵۲۸	۰/۴۷۲	۱	۰/۲۲۰
۱۹	ALL B-cell	۰/۶۱۸	۰/۳۸۲	۱	۰/۱۲۸
۲۰	ALL B-cell	۰/۷۳۱	۰/۲۶۹	۱	۰/۱۰۸
۲۱	AML	۰/۱۷۸	۰/۸۲۲	۲	۰/۳۵۸
۲۲	AML	۰/۱۶۵	۰/۸۳۵	۲	۰/۳۳۱
۲۳	AML	۰/۱۳۴	۰/۸۶۶	۲	۰/۳۲۹
۲۴	AML	۰/۱۲۱	۰/۸۷۹	۲	۰/۳۱۴
۲۵	AML	۰/۲۵۸	۰/۷۴۲	۲	۰/۲۸۶
۲۶	AML	۰/۱۳۸	۰/۸۶۲	۲	۰/۲۸۶
۲۷	AML	۰/۱۳۶	۰/۸۶۴	۲	۰/۲۶۵
۲۸	AML	۰/۲۶۶	۰/۷۳۴	۲	۰/۲۶۳
۲۹	AML	۰/۲۴۸	۰/۷۵۲	۲	۰/۲۵۶
۳۰	AML	۰/۱۴۵	۰/۸۵۵	۲	۰/۲۴۵
۳۱	AML	۰/۶۸۵	۰/۳۱۵	۱	۰/۲۰۲
۳۲	AML	۰/۱۱۷	۰/۸۸۳	۲	۰/۱۲۶
۳۳	AML	۰/۲۰۶	۰/۷۹۴	۲	۰/۱۱۳
۳۴	AML	۰/۱۴۹	۰/۸۵۱	۲	۰/۱۰۸

نمودار ۱ نمونه‌ها را در دو خوشه با روش خوشه‌بندی فازی نشان می‌دهد. با توجه به نمودار ۱ نتایج انتساب نمونه‌ها به دو خوشه در جدول۲ آمده است. با توجه به این جدول ویژگی روش خوشه‌بندی فازی در تشخیص افراد AML، ۹۰٪، حساسیت ۹۳٪ و صحت ۹۱٪ به دست آمد.

نمودار ۱ بیانگر عدم فشردگی خوشه‌ها می‌باشد، برای ارزیابی کارایی روش خوشه‌بندی فازی از نمودار سایه‌نما استفاده کردیم. برای رسم این نمودار ابتدا مقدارهای سایه‌نما را به دست آوردیم، سپس با توجه به مقادیر سایه‌نما، نمودار سایه‌نما را که در نمودار ۲ به دست آوردیم که ضریب سایه‌نما ۰/۳۱ به دست آمد، لذا ساختار خوشه‌ای ضعیف است و توصیه می‌شود از الگوریتم دیگری استفاده شود. برای بررسی بیشتر خوشه‌بندی فازی در سه خوشه نیز انجام گردید که نتیجه آن در نمودار ۳ آمده است که با توجه به آن خوشه اول و دوم

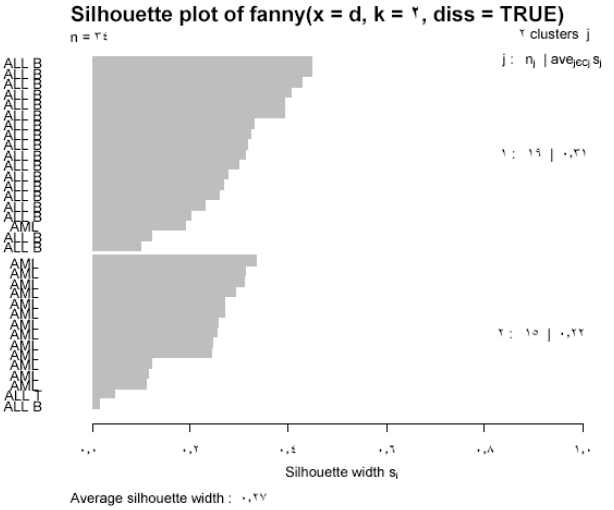
مربوط به ALL و خوشه سوم AML است. ضریب افراز دان ۰/۷۸۳ و شاخص غیرفازی ۰/۶۷۴ به دست آمد.



نمودار۱- نمایش نمونه‌ها در دو خوشه در روش خوشه‌بندی فازی

جدول۲- انتساب بیماران با روش خوشه‌بندی فازی به دو خوشه

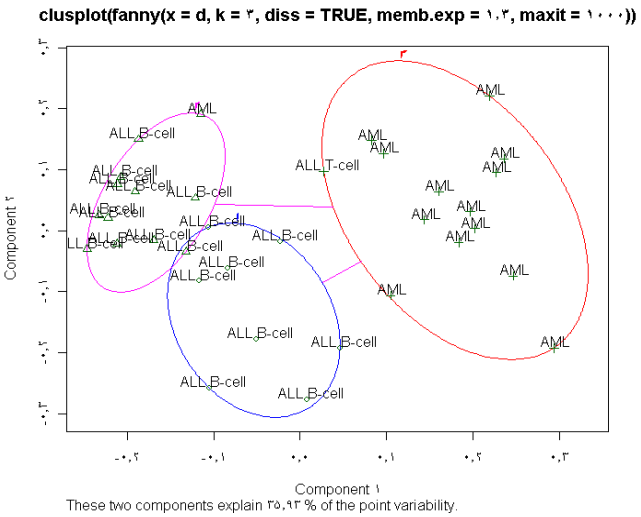
خوشه بندی / لوسمی	خوشه۱	خوشه۲	جمع
ALL	۱۸	۲	۲۰
AML	۱	۱۳	۱۴
حساسیت	ویژگی	صحت	
٪۹۳	٪۹۰	٪۹۱	



نمودار۲- نمودار سایه‌نما در دو خوشه به روش خوشه‌بندی فازی

بحث

با توجه به نتایج روش خوشه‌بندی فازی در انتساب نمونه‌ها به دو خوشه توانایی کامل دارد زیرا از ۲۰ نمونه ALL ۱۸ نمونه و از ۱۴ نمونه AML ۱۳ نمونه را به درستی اختصاص داده است.



نمودار ۳- نمایش نمونه‌ها در سه خوشه در روش خوشه‌بندی فازی

هر چقدر که تعداد خوشه‌ها را زیادتر کنیم، امکان یافتن زیرگروه‌های جدید افزایش می‌یابد، مثلاً در تعداد سه خوشه که در نمودار ۳ نشان داده شده است اعضای خوشه اول همگی ALL هستند. شاخص غیرفازی ۰/۷۴۸ به دست آمد که برتری روشهای افراز قطعی را نسبت به روش فازی در این داده‌ها نشان می‌دهد. ضریب سایه‌نما ۰/۳۱ این نظر را تأیید می‌کند، که در این داده‌های عملکرد روشهای دیگر خوشه‌بندی از فازی بهتر است، شاید دلیل آن عدم وجود موارد بینابین در داده‌ها است. در مواردی که در داده‌ها موارد بینابین وجود دارد روش فازی مناسبتر است چون امکان انتخاب را به پژوهشگر می‌دهد تا در مورد قرار گرفتن نمونه‌ها در خوشه‌ها تصمیم‌گیری کند. بر طبق نتایج خوشه‌بندی نمونه سی و یکم که بر اساس یافته‌های بالینی در گروه AML قرار دارد، در گروه ALL قرار گرفت، همچنین نمونه‌های دوم و هفدهم که بر اساس یافته‌های بالینی در گروه ALL قرار دارد طبق نتایج در گروه AML قرار گرفتند که از نظر بالینی می‌توانند قابل توجه باشند. نتایج روش خوشه‌بندی فازی با روشهای دیگر خوشه‌بندی که تاکنون بر روی این داده‌ها انجام شده است، همخوانی دارد (۱۵) ولی خوشه‌بندی فازی این برتری را دارد که محقق با بررسی ضرایب عضویت نمونه‌ها خصوصاً نمونه‌هایی که با یافته‌هایی بالینی مغایرت دارد تصمیم بهتری را بگیرد (۱۷و۱۶). همچنین با بررسی تعدادی از مطالعات بیان ژنی که از روش خوشه‌بندی فازی FCM استفاده کرده‌اند، می‌بینیم که از این روش می‌توان به منظور از بین بردن مشکل خوشه‌بندی اشتباه داده‌های ریزآرایه و مواردی که ساختار داده‌ها پیچیده است و خوشه‌بندی‌های رایج جواب نمی‌دهند، استفاده کرد(۷و۵). همچنین روش FCM در تحلیل‌های ریزآرایه در حالت مدلهای آمیخته

اعتمادپذیری و سهولت انجام بیشتری نسبت به مدلهای آمیخته دارد (۸). با توجه به این موارد استفاده از روش خوشه‌بندی فازی روز به روز افزونتر می‌گردد.

نتایج به دست آمده از ریزآرایه منجر به ایجاد دیدگاههای تازه‌ای در مورد نحوه شکل‌گیری، پیشرفت و پاسخ به درمان بیماران سرطانی گردیده است. با توجه به اینکه توالی کامل ژنوم انسانی در دسترس می‌باشد 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 شگرفی گردیده است. ترکیبی از روشهای ژنومیک و پروتئومیکس احتمالاً سبب پیشرفتهای عمیقتری در این زمینه خواهد شد (۱۸).

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 نمود. در واقع روشهای خوشه‌بندی برای مشخص کردن زیرگروه‌های احتمالی در مورد هر نوع سرطان دیگر را نیز به کار برد. روشهای کشف طبقات را می‌توان همچنین جهت تحقیق پیرامون مکانیزمهای اساسی که باعث تشخیص انواع سرطانها می‌شود به کار بست. به عنوان مثال، جهت مطالعه روند و نحوه بیان ژنها می‌توان سرطانهای متفاوتی را با هم در یک مجموعه واحد ترکیب نمود و بعد از حذف ژنهایی که با نوع بافتهای مورد نظر همبستگی بالایی دارند، نمونه‌ها را بر اساس ژنهای باقیمانده خوشه‌بندی کرد.

این روشها را می‌توان در کنار روشهای بالینی به کار 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 به کار بردن آنها در بیماران پرداخت. از طرفی روشهای 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 (۱۹).

فناوری تولید داده‌های ریزآرایه در ایران و بسیاری از کشورها هنوز مورد استفاده قرار نگرفته است و این نوع داده‌ها تنها در کشورهای معینی تولید می‌شود و پژوهشگرانی نظیر نویسندگان مقاله ناگزیر به استفاده از داده‌هایی هستند که توسط دیگران جمع‌آوری شده است، بنابراین کاستیهای احتمالی در صحت داده‌ها از محدودیت این تحقیق به شمار می‌آید.

نتیجه‌گیری

خوشه‌بندی فازی اطلاعات نسبتا قابل قبولی درباره ساختار داده‌ها فراهم می‌کند که با توجه به انطباق نتایج این روش با گروه‌بندی واقعی داده‌ها، از این روش آماری می‌توان در مواردی که اطلاع دقیقی درباره گروه‌بندی واقعی داده‌ها در دست نیست، استفاده کرد. به علاوه با بررسی نتایج خوشه‌بندی ممکن است زیرگروه‌هایی از نمونه‌ها را به نحوی متمایز کرد که برای انطباق آن با یافته‌های بالینی، پژوهشهای آزمایشگاهی یا بالینی جدیدی لازم باشد.

تشکر و قدردانی

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

REFERENCES

1. Beltrame F, Papadimitropoulos A, Porro I, Scaglione S, Schenone A, Torterolo L, et al. GEMMA-A Grid environment for microarray management and analysis in bone marrow stem cells experiments. Future Generation Computer Systems 2007;23(3):382-90.

2. Ho SY, Hsieh CH, Chen HM, Huang HL. Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. BioSystems 2006;85(3):165–76.

3. Satagopan JM, Panageas KS. A statistical perspective on gene expression data analysis. Stat Med 2003;22(3):481–99.

4. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863–8.

5. Woolf PJ, Wang Y .A fuzzy logic approach to analyzing gene expression data. Physiol Genomics 2000;3(1):9-15.

6. Inza I, Sierra B, Blanco R, Larranaga P. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. Journal of Intelligent and Fuzzy Systems 2002;12(1):25-34.

7. Dembele D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics 2003;19(8):973–80.

8. Asyali MH, Alci M. Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods. Bioinformatics 2005;21(5):644–49

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

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

11. 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(5439):531–7.

12. Cancer Program Data Sets. Available from: URL: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

13. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to cluster Analysis. Wiley: New York, 1990.

14. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational Application Mathematics 1987;20:53-65.

15. Vahedi M, Alavi Majd H, Mehrabi Y, Naghavi B. Gene expression data clustering and it's application in differential analysis of leukemia. Journal of Semnan University of Medical Sciences 2008;9(2):163-9. (Full Text in Persian)

16. Alavimajd H, Vahedi M, Mehrabi Y, Naghavi B. Clustering approach in DNA microarray analysis. Pejouhesh dar pezeshki 2007;31(1):19-26. (Full Text in Persian)

17. Vahedi M. Application of clustering methods in gene expression data [Dissertation]. Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran; 2006.

18. Korenberg MJ. Microarray Data Analysis: Methods and Applications (Methods in Molecular Biology). New Jersey: Humana Press; 2007.

19. Knudsen S. Cancer Diagnostics with DNA Microarrays. New Jersey: Wiley; 2006.