دانشور
پزشکی

دوماهنامه علمی–پژوهشی
دانشگاه شاهد
سال هیجدهم– شماره ۹۳
تیر ۱۳۹۰

به کارگیری خوشه‌بندی دوبعدی با روش «زیرماتریس‌های با میانگین- درایه‌های بزرگ» در داده‌های بیان ژنی حاصل از ریزآرایه‌های DNA

نویسندگان: حمید علوی‌مجد[1*]، شیما یونس‌پور[2]، فرید زایری[3]، مصطفی رضایی طاویرانی[4]

۱. دانشیار گروه آمار زیستی دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران
۲. دانشجوی کارشناسی ارشد آمار زیستی دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران
۳. استادیار گروه آمار زیستی دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران
٤. دانشیار گروه علوم پایه دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران
* نویسنده مسئول: حمید علوی مجد **E-mail: alavimajd@gmail.com**

دریافت: ۱۳۸۹/۱۱/۲۵
آخرین اصلاح‌ها: ۱۳۹۰/۲/۱۹
پذیرش: ۱۳۹۰/۲/۲۴

چکیده

مقدمه و هدف: در سال‌های اخیر، فناوری ریزآرایه‌ی DNA نقش اساسی در تحقیقات ژنومی داشته‌است. با استفاده از این فناوری که امکان آنالیز هم‌زمان سطوح بیان هزاران ژن را در شرایط مختلف فراهم‌آورده‌است، به حجم انبوهی از داده‌ها دست‌می‌یابیم. روش‌های کلاسیک خوشه‌بندی نظیر روش‌های سلسله‌مراتبی و غیرسلسله‌مراتبی، روش‌هایی مناسب برای تحلیل این‌گونه داده‌ها هستند اما محدودیت‌هایی نیز دارند. در این روش‌ها فرض بر آن است که یک ژن یا یک شرایط آزمایشی را تنها می‌توان به یک خوشه منتسب‌کرد و یک ژن، متعلق به گروهی از ژن‌هاست که با هم، تحت همه شرایط آزمایشی تنظیم می‌شوند. بنابراین به‌منظور رفع این کاستی‌ها از روش‌های خوشه‌بندی دوبعدی استفاده‌می‌شود. هدف از این پژوهش، بررسی کارآیی یک روش خوشه‌بندی دوبعدی در تحلیل داده‌های بیان ژنی مخمر است.

مواد و روش‌ها: در این پژوهش، داده‌های بیان ژنی مخمرِ Saccharomyces cerevisiae گسچ و همکاران (۲۰۰۰) با استفاده از روش خوشه‌بندی دوبعدی (Large Average Submatrices; LAS) تحلیل‌شده‌اند. مجموعه داده‌ها، ۱۷۳ شرایط آزمایشی مختلف و مجموعه‌ای از ۲۹۹۳ ژن را دربرگرفته و برای تحلیل داده‌ها از نرم‌افزارهای LAS، JMP و GOAL استفاده‌شده‌است.

نتایج: نتایج نشان‌داد که روش LAS قادر است خوشه‌های دوبعدی مناسبی از دیدگاه آماری و زیست‌شناسی تولیدکند.

نتیجه‌گیری: این مطالعه نشان‌می‌دهد که می‌توان با استفاده از روش LAS، زیرمجموعه‌هایی از ژن‌ها را با الگوهای بیان مشابه در زیرمجموعه‌ای از شرایط آزمایشی شناسایی‌کرد که از نظر زیست‌شناسی معنی‌دارند.

واژگان کلیدی: خوشه‌بندی دوبعدی (Biclustering)، داده‌های بیان ژنی، ریزآرایه DNA، هستی‌شناسی ژنی (gene ontology)، زیرماتریس‌های با میانگین-درایه‌های بزرگ (Large Average Submatrices)

مقدمه

در سال‌های اخیر، فناوری ریزآرایه (microarray) DNA، نقشی اساسی در تحقیقات علوم زیستی داشته و تحولی بزرگ در زمینه علوم پزشکی به‌وجودآورده و از-این‌رو مورد توجه بسیاری از پژوهشگران این رشته‌ها قرارگرفته‌است. ریزآرایه DNA عبارت است از مجموعه‌ای از نقاط میکروسکوپی DNA که به سطحی جامد، مانند شیشه، پلاستیک یا تراشه سیلیکون متصل شده‌اند و یک آرایه را تشکیل‌می‌دهند (۱). در گذشته، محققان تنها به اندازه‌گیری تعداد به‌نسبت کمی از ژن‌ها در هر بار آزمایش قادر بودند؛ امروزه با ظهور فناوری ریزآرایه، امکان اندازه‌گیری سطوح بیان تعداد زیادی از ژن‌ها به‌طور هم‌زمان در یک آزمایش تکی و استفاده از این روش در بازه وسیعی از تحلیل‌های ژنومیک، نظیر شناسایی ژن‌ها، اکتشاف داروها، تشخیص‌های کلینیکی و کشف زیرنوع‌های (subtypes) بیماری‌ها فراهم‌آمده-است (۲). الگوی بیان ژن یک سلول یا یک بافت، ساختار و عملکرد آن را مشخص‌می‌کند و داده‌های سطوح بیان ژن‌ها، اطلاعاتی مفید را در زمینه شبکه‌های بیولوژیک و فهم فرایندهای سلولی فراهم‌می‌سازد (۲).

از کاربردهای تحقیقاتی معمول ریزآرایه می‌توان به مواردی از قبیل تعیین چگونگی بیان ژن‌ها در دو شرایط مختلف، تعیین کارکرد ژن‌های ویژه در یک وضعیت به-خصوص، مانند شرایط غذایی، دمایی یا شیمیایی ویژه، تعیین چگونگی تاثیر بیان هر ژن روی بیان ژن‌های دیگر در همان شبکه ژنتیکی، پیش‌بینی پیشرفت بیماری و تعیین میزان موفقیت درمان و دستیابی به درمان‌های ویژه هر بیمار بر اساس نتایج حاصل از بیان ژن‌ها اشاره‌کرد. درمجموع با استفاده فناوری ریزآرایه به دامنهٔ وسیعی از پرسش‌های زیست‌شناسی در مورد تعداد بسیاری از ژن‌ها و در مواردی، کل ژن‌های یک ژنوم پاسخ‌داده-خواهدشد (۱).

داده‌های بیان ژنی حاصل از ریزآرایه‌ی DNA، اغلب به‌صورت ماتریسی از سطوح بیان ژن‌ها، تحت شرایط آزمایشی مختلف نشان‌داده‌می‌شود؛ سطرهای این ماتریس نشان‌دهنده‌ی ژن‌ها و ستون‌های آن بیانگر شرایط (یا نمونه‌های) آزمایشی است. ممکن است نمونه‌ها مربوط به زمان‌ها، شرایط محیطی یا حتی افراد گوناگون باشد.

از طریق ریزآرایه، به حجم بسیار زیادی از داده‌ها دست‌می‌یابیم و نیاز به روش‌ها و ابزارهای دقیق و توانمند برای تحلیل این داده‌ها به‌طور کامل محسوس است. خوشه‌بندی (clustering) از جمله روش‌های تحلیلی است که ما را در تفسیر داده‌های بیان ژن، با یافتن گروه‌هایی از ژن‌ها با الگوهای بیان مشابه، یاری-می‌رساند (۳). خوشه‌بندی درواقع تقسیم‌بندی یک جمعیت ناهمگون به تعدادی از زیرمجموعه‌های همگون است که به آنها خوشه اطلاق‌می‌شود. در این روش به دنبال یافتن گروه‌هایی هستیم که با یکدیگر بسیار متفاوتند، ولی اعضای این گروه‌ها بسیار به هم شبیه هستند (٤)؛ اما با وجود توانمندی روش‌های کلاسیک خوشه‌بندی از قبیل روش‌های خوشه‌بندی سلسله‌مراتبی و k-میانگین در تحلیل داده‌های ریزآرایه، این روش‌ها محدودیت‌هایی نیز دارند (۵)؛ نخست اینکه روش‌های کلاسیک خوشه‌بندی بر این فرض استوارند که ژن‌های مرتبط، رفتاری مشابه در همه شرایط اندازه‌گیری‌شده دارند، این فرض وقتی منطقی است که مجموعهٔ داده‌ها، تعداد کمی از شرایط در یک آزمایش ساده را دربرگیرد، درحالی‌که برای مجموعه داده‌های بزرگ‌تر که صدها شرایط ناهمگون از تعداد زیادی از آزمایش‌ها را شامل می‌شوند، این فرض برقرار نیست و ازاین‌رو، استفاده از روش‌های کلاسیک خوشه‌بندی، منطقی به‌نظرنمی‌رسد. درحقیقت، درک کلی ما از فرایند سلولی موجب می‌شود که انتظارداشته‌باشیم زیرمجموعه‌های ژن‌ها، تنها طی شرایط آزمایشی خاصی بیان و تنظیم شوند و در شرایط دیگر تقریبا مستقل عمل‌کنند؛ همچنین در روش‌های کلاسیک خوشه‌بندی، ژن‌ها به مجموعه‌هایی دوبه‌دو مجزا از هم دسته‌بندی‌می‌شوند که بیانگر پیوند هر ژن با یک فرایند یا کارکرد زیستی ساده است و این شاید ساده‌انگاری بیش‌از حد سیستم زیستی باشد (۳).

به‌منظور رفع کاستی‌های روش‌های کلاسیک خوشه‌بندی، از روش‌های خوشه‌بندی دوبعدی (biclustering) استفاده‌می‌شود؛ در این روش‌ها، زیرمجموعه‌ای از ژن‌ها با الگوهای بیان مشابه در زیرمجموعه‌ای از شرایط جستجومی‌شوند؛ یعنی، زیرماتریس‌های همگنی (خوشه‌های دوبعدی) در ماتریس داده‌های بیان ژنی یافته‌می‌شوند و الگوهای موضعی (local patterns) از این داده‌ها به‌دست‌می‌آیند (۳).

هارتیگان، خوشه‌بندی دوبعدی را در سال ۱۹۷۵ مطرح‌کرد (٦). *چنگ و چرچ* در سال ۲۰۰۰، اولین افرادی بودند که این روش را در داده‌های بیان ژنی به‌کاربردند؛ آنها یک خوشه‌ی دوبعدی را در جایگاه یک زیرماتریس یکنواخت با میانگین مربع‌های مانده‌های (mean squared residue) کوچک تعریف‌کردند و یک رویکرد گریدی (greedy) برای یافتن خوشه‌های دوبعدی به‌کاربردند (۷). *لزرونی و اوون*، نظریه یک مدل شطرنجی (plaid) را مطرح‌کردند؛ در این روش، ماتریس ورودی، یک تابع خطی از متغیرهای مربوط به خوشه‌بندی دوبعدی درنظرگرفته‌می‌شود. آنها نشان‌دادند که چگونه با استفاده از یک فرایند ماکسیمم‌سازی تکراری می‌توان یک مدل را برآوردکرد (۸). *بن-دور و همکاران* در سال ۲۰۰۲ یک خوشه دوبعدی را به عنوان یک زیرماتریس حافظ ترتیب (Order preserving submatrix) تعریف‌کردند (۹). در سال ۲۰۰٤ یک بررسی دربارهٔ چند روش خوشه‌بندی دوبعدی انجام‌شد (۱۰). *پرلیک و همکاران* در سال ۲۰۰٦، مقایسه‌ای سیستماتیک دربارهٔ تعدادی از روش‌های خوشه‌بندی دوبعدی انجام‌دادند (۱۱). در سال‌های ۲۰۰۷-۲۰۰٦ تعدیل‌های مختلفی برای مدل شطرنجی در مقالات پیشنهادشده‌است. *گو و لیو* در سال ۲۰۰۸، روش خوشه‌بندی دوبعدی بیزی را در داده‌های بیان ژنی به‌کاربردند (۵). *شابالین و همکاران* در سال ۲۰۰۹ روش آماری خوشه‌بندی دوبعدی «زیرماتریس‌های با میانگین-درایه‌های بزرگ» (Large Average Submatrices; LAS) را معرفی‌کردند؛ این روش، زیرماتریس‌هایی با میانگین-درایه‌های بزرگ را در ماتریس داده‌های با مقادیر حقیقی جستجومی‌کند (۱۲).

هدف از این پژوهش، بررسی کارآیی روش خوشه‌بندی دوبعدی LAS در تحلیل داده‌های بیان ژنی مخمرِ Saccharomyces cerevisiaeاست. این مخمر، یکی از پراستفاده‌ترین مدل ارگانیسم‌ها در علم به شمار می‌آید و به دلیل کاربردهای بسیار آن در صنعت، جایگاهی مهم را به خود اختصاص داده است؛ از کاربردهای این مخمر می‌توان به تخمیر آبجو و نان و تهیه اتانول اشاره-کرد؛ علاوه بر این، مخمرها از نظر ساختار، به نسبت، مشابه سلول‌های انسان و برخلاف باکتری‌ها هر دو موجوداتی، یوکاریوتی هستند؛ به علاوه، بسیاری از پروتئین‌های مهم در بدن انسان در ابتدا با مطالعه تشابه‌شان در مخمر کشف شده‌اند. این پروتئین‌ها پروتئین‌های چرخه سلولی، پروتئین‌های پیام‌رسانی و آنزیم‌های پردازش‌کنندهٔ پروتئین‌ها را در برمی‌گیرند؛ همچنین جهش کوچک (petite mutation) در مخمر یاد شده، یکی از موارد مورد علاقه تحقیق در زیست‌شناسی مولکولی محسوب می‌شود (۱۳).

مواد و روش‌ها

داده‌های پژوهش

در این پژوهش، از داده‌های مخمرِ Saccharomyces cerevisiae که *گسچ و همکاران* در سال ۲۰۰۰ انتشار-دادند، استفاده‌شده‌است (١٤). مجموعهٔ داده‌ها ۱۷۳ شرایط آزمایشی مختلف و ۲۹۹۳ ژن را دربرمی‌گیرد. شرایط آزمایشی استفاده‌شده در داده‌ها (مواردی نظیر شوک گرمایی، کمبود نیتروژن و ...) هستند (۱۵). توضیح بیشتر درباره‌ی داده‌های پژوهش همراه با فهرست مقالاتی که از این داده‌ها استفاده‌کرده‌اند، در فهرست منابع مقالهٔ حاضر آمده‌است (۱۰)؛ این داده‌ها از طریق اینترنت در اختیار عموم قراردارند (١٦).

الگوریتم LAS

در این تحقیق، از روش خوشه‌بندی دوبعدی LAS، برای تحلیل داده‌های بیان ژنی مخمر موردنظر، استفاده-

شده‌است. این روش، رویکردی مبتنی بر معنی‌داری، برای خوشه‌بندی دوبعدی داده‌هایی با مقادیر حقیقی را ارائه‌می‌دهد.

فرض کنیم:

$X=\{x_{ij}: i\in\{1,2,...,m\}\ , j\in\{1,2,...,n\}\}$ ماتریس داده‌های مشاهده‌شده باشد؛ یک زیرماتریس از X، مجموعه‌ای اندیس‌گذاری‌شده از درایه‌ها به‌صورت $U=\{x_{ij}:\ i\in A\ , j\in B\}$ است به‌طوری‌که $A\subseteq\{1,2,...,m\}$ و $B\subseteq\{1,2,...,n\}$.

الگوریتم **LAS** بر پایه یک مدل زیرماتریس جمع‌پذیر است که طی آن، ماتریس داده‌های X به صورت مجموع K زیرماتریس ثابت و به‌طور بالقوه متداخل (**overlapping**) به‌علاوهٔ خطا (**noise**) بیان‌می‌شود؛ به‌طور دقیق‌تر، مدل به شکل زیر تعریف‌می‌شود.

$$i\in\{1,2,...,m\}\ , j\in\{1,2,...,n\} \qquad (۱)$$

که در آن به طوری که $A_k\subseteq\{1,2,...,m\}$ و $B_k\subseteq\{1,2,...,n\}$ مجموعه‌های سطری و ستونی k اُمین زیرماتریس، $\alpha_k\in\mathbb{R}$ سطح k اُمین زیرماتریس و $\{\varepsilon_{ij}\}$ متغیرهای تصادفی مستقل $N(0,1)$ هستند. در معادله ۱، $I(.)$، تابعی نشانگر است که وقتی شرایط داخل پرانتز برقرار باشد، مقدار آن برابر ۱ می‌شود؛ وقتی $K=0$ باشد، مدل ۱ به‌صورت مدل ۲ (مدل فرضی $H_0: K=0$) کاهش‌می‌یابد که در شرایط مدل ۲، X یک ماتریس تصادفی گاوسی $m\times n$ است.

$$\{x_{ij}: i\in\{1,2,...,m\}\ , j\in\{1,2,...,n\}\} \qquad (۲)$$

$x_{ij}\overset{i.i.d}{\sim}N(0,1)$ که در آن است.

$$S(U)=-\log\left[\binom{m}{k}\binom{n}{l}\Phi\left(-\tau\sqrt{kl}\right)\right]$$

مدل ۲ به‌طور طبیعی به یک تابع امتیاز مبتنی بر معنی‌داری برای زیرماتریس‌ها منجرمی‌شود. به‌ویژه، امتیاز اختصاص‌یافته به زیرماتریس $U_{k\times l}$ از ماتریس داده‌های X با $Avg(U)=\tau>0$ به‌صورت زیر تعریف می‌شود:

$$x_{ij}=\sum_{k=1}^{K}\alpha_k I\left(i\in A_k, j\in B_k\right)+\varepsilon_{ij}$$

که منظور از $Avg(U)$، مقدار میانگین درایه‌های زیرماتریس U است.

از دیدگاه آزمون آماری، عبارت داخل کروشه، یک p-مقدار اصلاح‌شدهٔ بونفرونی مربوط به مدل ۲ و تابع آزمون $Avg(U)$ است. تابع امتیاز $S(.)$، انحراف‌ها از مدل ۲ را با درنظرگرفتن ابعاد و مقدار میانگین یک زیرماتریس، اندازه‌گیری‌می‌کند و یک معیار سادهٔ یک‌بعدی را فراهم‌می‌سازد؛ با این معیار، می‌توان زیرماتریس‌های مشاهده-شده با اندازه‌ها و مقادیر میانگین (درایه‌های) مختلف را مقایسه و رتبه‌بندی‌کرد؛ در این تابع، در میان زیرماتریس‌هایی با ابعاد یکسان، زیرماتریس‌هایی با میانگین بزرگ‌تر، مورد توجه قرارمی‌گیرند.

تابع امتیاز **LAS** بر اساس تابع توزیع تجمعی نرمال بوده، به انحراف‌های ناشی از فرض نرمالیتی که ناشی از دنباله‌های سنگین (**heavy tail**) در توزیع تجربی مقادیر بیان ژنی است، حساس می‌باشد. نقاط دورافتاده نیز می‌توانند به زیرماتریس‌هایی منجرشوند که با وجود معنی‌داری قابل توجه، تعداد ژن‌ها یا تعداد شرایط بسیار کمی را دربرگیرند.

برای اولین گام در الگوریتم، نمودار **Q-Q** استاندارد توزیع تجربی درایه‌های ماتریس داده‌هایی که به‌صورت ستونی استاندارد شده‌اند در مقابل تابع توزیع تجمعی نرمال استاندارد درنظرگرفته‌می‌شود. در صورت مشاهدهٔ دنباله‌های سنگین در توزیع تجربی داده‌های بیان ژنی از تبدیل زیر برای هر درایه از ماتریس داده‌ها استفاده-می‌شود.

$$f(x)=sign(x)\log\left(1+|x|\right)$$

شیوهٔ عملکرد الگوریتم **LAS**، در یافتن زیرماتریس‌ها (خوشه‌های دوبعدی)، به‌طور کامل در مقاله منتشرشدهٔ *شابالین و همکاران*، توضیح‌داده‌شده‌است (۱۲).

در بسیاری از روش‌های خوشه‌بندی دوبعدی نیاز است که کاربر، تعدادی پارامتر عملیات را تعیین‌کند و در بیشتر موارد، عملکرد بهینه‌ی روش، به انتخاب دقیق پارامترها نیازمند است؛ علاوه‌براین، برای الگوریتم‌های دقیق (**exact**)، تغییر اندک پارامترها می‌تواند به تغییرهایی

قابل توجه در جهت اندازه و قابل تفسیرشدن خروجی منجرشود. تنها پارامترهای عملیاتی در الگوریتم LAS عبارت‌اند از: ۱) تعداد دفعاتی که رویکرد اصلی جستجو در هر حلقهٔ اصلی الگوریتم اجرامی‌شود و ۲) معیار توقف. کم‌بودن تعداد پارامترها، یک خصوصیت مهم الگوریتم LAS است که کاربرد این روش را در مسایل علمی به‌نسبت ساده‌می‌سازد (۱۲).

برای یافتن خوشه‌های دوبعدی و تحلیل‌های آماری این پژوهش، از نرم‌افزارهای LAS و JMP استفاده‌شده.

روش اعتبارسنجی نتایج حاصل از خوشه‌بندی دوبعدی

در این پژوهش، اطلاعات موجود زیست‌شناسی، به‌منظور اعتبارسنجی الگوریتم خوشه‌بندی دوبعدی LAS (در یافتن خوشه‌های دوبعدی) به‌کارگرفته‌شده‌است. برای تعیین معنی‌داری خوشه‌های دوبعدی از نظر زیستی، از پایگاه داده‌های تفسیر هستی‌شناسی ژنی (Gene Ontology; GO) استفاده‌شده‌است. هستی‌شناسی ژنی، واژگان کنترل شده‌ای برای توصیف عملکرد مولکولی (Molecular Function; MF)، فرایند زیستی (Biological Process; BP) و جایگاه سلولی (Cellular Component; CC) فرآورده‌های ژنی، فراهم‌می‌سازد (۱۷و۱۹).

توجه اصلی پژوهش در این مطلب است که «آیا خوشه‌های دوبعدی کشف‌شده با روش LAS، ژن‌هایی با کارکرد یکسان را شامل می‌شود؟».

به‌منظور ارزیابی توانایی روش LAS، در یافتن خوشه‌های دوبعدی با ویژگی «کارکرد یکسان ژن‌ها در خوشه‌ها»، بایدبررسی‌شود که آیا مجموعه ژن‌های به‌دست‌آمده با این روش، غنی‌سازی معنی‌داری (significant enrichment) را بر اساس تفسیر GO خاص، نشان‌می‌دهند یا خیر؟ (۲۰و۲۱).

در این مقاله، درصد خوشه‌های دوبعدی که در یک یا چند تفسیر هستی‌شناسی ژنی (GO)، بیش نشان‌داده (overrepresented) شده‌اند، بر اساس هریک از دسته‌های کارکردی BP و MF و CC در سطوح معنی‌داری مختلف، محاسبه شده‌است (۲۱)؛ یعنی درصد خوشه‌های دوبعدی غنی‌شدهٔ معنی‌دار، بر اساس هریک از دسته‌های کارکردی در سطوح معنی‌داری مختلف، به‌دست‌آمده‌است.

در این پژوهش، نرم‌افزار (Gene Ontology AnaLyzer; GOAL) به‌منظور ارزیابی خوشه‌های دوبعدی به‌دست‌آمده، به‌کاررفته‌است. این نرم‌افزار، همه عبارت‌های GO تفسیر شده و عبارت‌های GO والد آنها (parental GO terms) را برای مجموعه ژن‌های ورودی (ژن‌های موجود در خوشهٔ دوبعدی موردنظر) و مجموعه ژن‌های مرجع (ژن‌های موجود در توزیع پیش‌زمینه) تعیین‌می‌کند؛ سپس، تعداد دفعات وجود هریک از عبارت‌های GO را برای ژن‌های موجود در خوشهٔ دوبعدی موردنظر و برای ژن‌های مجموعه مرجع، محاسبه‌می‌کند. از آزمون دقیق فیشر (Fisher's exact test) برای تعیین معنی‌داری اختلاف مشاهده‌شده استفاده‌می‌شود؛ با این کار، یک p–مقدار (p-value) برای هر عبارت GO به‌دست‌می‌آید. برای به‌دست‌آوردن توان آزمون بیشتر، (با کاربرد یکی از سه روش موجود در نرم‌افزار GOAL)، p–مقدار اصلاح‌می‌شود. p–مقدار اصلاح‌شده، بیانگر این مطلب است که ژن‌های موجود در خوشه دوبعدی موردنظر، تا چه حد با دسته‌های GOمختلف، هماهنگی دارند. p–مقدار کوچک‌تر (نزدیک به صفر)، نشان‌دهندهٔ هماهنگی بیشتر، میان ژن‌های موجود در خوشه دوبعدی و دسته‌های GO مختلف است (۲۲).

مقداردهی پارامترهای مدل

با توجه به مشاهده دنباله‌های سنگین در توزیع تجربی داده‌ها، از تبدیل مربوط به نرمال‌سازی توزیع داده‌ها استفاده‌شد. با انجام این پیش‌پردازش، مجموعه داده‌ها برای خوشه‌بندی دوبعدی با نرم‌افزار LAS، آماده گردید.

مقداردهی پارامترهای مربوط به معیارهای توقف الگوریتم به این ترتیب درنظرگرفته‌شد که مقدار ۳۰ برای

تعداد خوشه دوبعدی (مثبت/منفی) موردنظر و مقدار ۵۰ برای نقطه برش تابع امتیاز. یعنی، وقتی که $S(U^*)$ کمتر از مقدار آستانه‌ای ۵۰ شود یا هنگامی که الگوریتم، موفق به یافتن ۳۰ خوشه‌ی دوبعدی (مثبت/منفی) شود، جستجوی خوشه‌های دوبعدی، پایان‌می‌یابد. قابل ذکر است که مقادیر آستانه‌ای و تعداد خوشه‌های دوبعدی بر مبنای روش تکراری مورد استفاده در مقالات مشابه بهینه‌شد.

شابالین و همکاران با انجام آزمایش‌هایی در داده‌های واقعی دریافتند که ۱۰۰۰ بار تکرار حلقه‌ی اصلی رویکرد جستجو، برای تضمین پایداری عملکرد الگوریتم کافی است (۱۲)؛ ازاین‌رو، در اجرای الگوریتم، مقدار این پارامتر، برابر ۱۰۰۰ درنظرگرفته‌شد.

نتایج

در ماتریس داده‌های مورد بررسی، دامنهٔ میانگین‌های ستون‌ها (۰/۱۳, ۰/۵۸-) و دامنهٔ انحراف معیارهای ستون‌ها (۱/۹۵, ۰/۲۵) مشاهده‌شد.

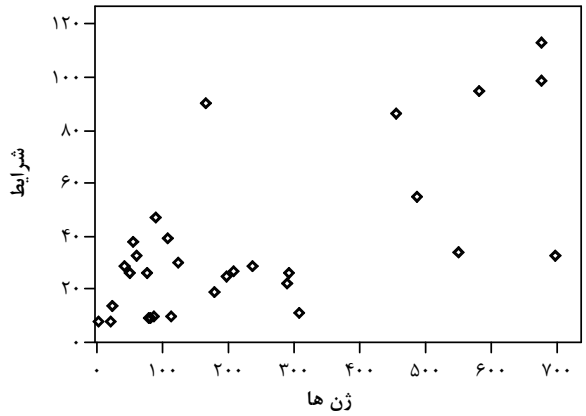
در مجموعه داده‌های مورد بررسی، سی خوشه دوبعدی با الگوریتم LAS، شناسایی‌شد. در شکل ۱، ابعاد سطری و ستونی خوشه‌های دوبعدی تولیدشده با الگوریتم، یعنی ژن‌ها و شرایط، نشان‌داده‌شده‌است؛ در این نمودار پراکنش، دامنهٔ وسیعی از اندازه‌های خوشه‌های دوبعدی با مینیمم اندازه ۴×۸ (تعداد شرایط × تعداد ژن‌ها) و ماکسیمم اندازه ۱۱۳×۶۷۶، مشاهده‌می‌شود.

در جدول ۱، درصد خوشه‌های دوبعدی غنی شده معنی‌دار، بر اساس هریک از دسته‌های کارکردی BP و MF و CC در سطوح معنی‌داری مختلف، آورده‌شده‌است. با توجه به این جدول، بیشترین غنی‌سازی، در دسته »فرآیند زیستی« مشاهده‌می‌شود. به‌طور نمونه، درصد خوشه‌های دوبعدی غنی‌شده بر اساس این دسته در سطح معنی‌داری ۵٪، ۹۳/۳۳٪ است؛ یعنی، تعداد ۲۸ خوشه از ۳۰ خوشهٔ دوبعدی تولیدشدهٔ LAS، دارای دست‌کم یک عبارت غنی‌شده معنی‌دار (در سطح معنی‌داری ۵٪) است.

جدول ۲، معنی‌دارترین عبارت‌های GO مورد استفاده برای توصیف ۶۱ ژن موجود در یکی از خوشه‌های دوبعدی (خوشه دوبعدی ۱۳) را، در دسته کارکردی «جایگاه سلولی» نشان‌می‌دهد. p-valueهای اصلاح‌شده در ستون‌های ۵ و ۶ این جدول آورده‌شده‌است. به‌طور نمونه، از ۲۹۹۳ ژن موجود در مجموعه مرجع، ۷۵ ژن به عبارت GO:0044455 تفسیرشده‌اند. از ۶۱ ژن موجود در خوشه دوبعدی موردنظر، ۲۱ ژن نیز به همین عبارت تفسیرشده‌اند. با استفاده از توزیع فوق هندسی، یک p-مقدار برای این عبارت به‌دست‌می‌آید ($p=۶/۹۷\times۱۰^{-۲۰}$). این مطلب به این معنی است که خوشه دوبعدی شامل ژن‌هایی است که از نظر زیستی تقریبا به هم شبیه‌اند و روش خوشه‌بندی دوبعدی به شناسایی خوشه‌های دوبعدی‌ای قادر است که از نظر زیستی معنی‌دارند.

بحث و نتیجه‌گیری

در این پژوهش، الگوریتم خوشه‌بندی دوبعدی LAS، در تحلیل داده‌های بیان ژنی مخمر Saccharomyces cerevisiae، به‌کارگرفته‌شد. تنها پارامترهای عملیاتی مورد نیاز در این الگوریتم، معیار توقف الگوریتم و تعداد جستجوهای تصادفی انجام‌شده، در شناسایی هر خوشه دوبعدی، بود. کم‌بودن تعداد پارامترهای عملیاتی، از خصوصیات مهم این الگوریتم به‌شمارمی‌رود.



نمودار ۱. نمودار پراکنش تعداد شرایط آزمایشی در برابر تعداد ژن‌های خوشه‌های دوبعدی به‌دست‌آمده به روش LAS

جدول ۱. درصد خوشه‌های دوبعدی غنی شدهٔ معنی‌دار، در سطوح مختلف معنی‌داری بر اساس دسته‌های کارکردی

p-مقدار اصلاح‌شده به روش بونفرونی		دسته‌های کارکردی	
	فرآیند زیستی (BP)	کارکرد مولوکولی (MF)	جایگاه سلولی (CC)
$p < ۰/۰۵$	۹۳/۳۳٪	۷۳/۳۳٪	۷۶/۶۷٪
$p < ۰/۰۱$	۹۰/۰۰٪	۷۰/۰۰٪	۶۶/۶۷٪
$p < ۰/۰۰۵$	۸۶/۶۷٪	۶۶/۶۷٪	۶۶/۶۷٪
$p < ۰/۰۰۱$	۸۳/۳۳٪	۶۶/۶۷٪	۶۳/۳۳٪
$p < ۱۰^{-۴}$	۷۳/۳۳٪	۵۶/۶۷٪	۶۰/۰۰٪
$p < ۱۰^{-۵}$	۷۰/۰۰٪	۵۶/۶۷٪	۵۳/۳۳٪

جدول ۲. معنی‌دارترین عبارت‌های GO (دسته »جایگاه سلولی«) برای یک خوشه دوبعدی به‌دست‌آمده به روش LAS

GO ID	تفسیر " جایگاه سلولی"	تعداد دفعات وجود عبارت GO در مجموعه مرجع	تعداد دفعات وجود عبارت GO در خوشه‌ی دوبعدی موردنظر	p-مقدار	p-مقدار اصلاح شده به روش بونفرونی
GO:0070469	respiratory chain	۲٤	۱۳ (٥٤/۱۷٪)	$۵/۷۴\times۱۰^{-۱۷}$	$۱/۱۹\times۱۰^{-۱۴}$
GO:0005746	mitochondrial respiratory chain	۲۱	۱۳ (٦۱/۹۰٪)	$۴/۸۹\times۱۰^{-۱۸}$	$۱/۶۹\times۱۰^{-۱۵}$
GO:0044455	mitochondrial membrane part	۷۵	۲۱ (۲۸/۰۰٪)	$۶/۹۷\times۱۰^{-۲۰}$	$۷/۲۴\times۱۰^{-۱۷}$
GO:0019866	organelle inner membrane	۱۳۸	۲۳ (۱٦/٦۷٪)	$۲/۵۶\times۱۰^{-۱۶}$	$۴/۴۲\times۱۰^{-۱۴}$
GO:0005740	mitochondrial envelope	۱۹۷	۲٥ (۱۲/٦۹٪)	$۷/۱۸\times۱۰^{-۱۵}$	$۸/۲۶\times۱۰^{-۱۳}$
GO:0031966	mitochondrial membrane	۱۹۲	۲٥ (۱۳/۰۲٪)	$۳/۸۵\times۱۰^{-۱۵}$	$۴/۹۶\times۱۰^{-۱۳}$
GO:0005743	mitochondrial inner membrane	۱۳۳	۲٥ (۱۸/۸۰٪)	$۳/۷۷\times۱۰^{-۱۹}$	$۱/۹۶\times۱۰^{-۱۶}$
GO:0031967	organelle envelope	۲۸٤	۲٥ (۸/۸۰٪)	$۳/۷۳\times۱۰^{-۱۱}$	$۲/۷۶\times۱۰^{-۹}$
GO:0031090	organelle membrane	٥٤۰	۲۷ (٥/۰۰٪)	$۱/۵۴\times۱۰^{-۶}$	$۶/۳۰\times۱۰^{-۵}$
GO:0044425	membrane part	۷۱۳	۳۲ (٤/٤۹٪)	$۹/۷۵\times۱۰^{-۷}$	$۴/۱۹\times۱۰^{-۵}$
GO:0044429	mitochondrial part	۳٤٥	۳۲ (۹/۲۸٪)	$۲/۳۷\times۱۰^{-۱۵}$	$۳/۵۱\times۱۰^{-۱۳}$
GO:0005739	mitochondrion	٥۹۹	٤۳ (۷/۱۸٪)	$۸/۷۳\times۱۰^{-۱۸}$	$۲/۲۶\times۱۰^{-۱۵}$
GO:0044444	cytoplasmic part	۱٤۹٥	٥۰ (۳/۳٤٪)	$۱/۷۱\times۱۰^{-۷}$	$۸/۰۲\times۱۰^{-۶}$
GO:0005737	cytoplasm	۱۹۲۹	٥٥ (۲/۸٥٪)	$۳/۷۰\times۱۰^{-۶}$	$۱/۳۷\times۱۰^{-۴}$

به‌منظور ارزیابی روش LAS (در یافتن خوشه‌های دوبعدی)، از اندازه‌های اعتبارسنجی آماری و زیستی استفاده‌شد. با توجه به نتایج آماری، این الگوریتم قادر بوده‌است که خوشه‌های دوبعدی با دامنهٔ وسیعی از اندازه‌های ژن‌ها و شرایط را بیابد؛ یعنی این الگوریتم، می‌توانسته پیوندهای سطری- ستونی را در بازه‌ای از اندازه‌های مختلف کشف‌کند.

با توجه به نتایج اعتبارسنجی زیستی، روش خوشه‌بندی دوبعدی LAS، درصدی به‌نسبت بالا را از خوشه‌های دوبعدی غنی‌شده معنی‌دار در سطوح معنی‌داری مختلف به‌دست‌می‌دهد (از ٪۷۳/۳۳ تا ٪۹۳/۳۳ در دسته BP). این مطلب نشان‌دهنده توانایی این روش در یافتن خوشه‌های دوبعدی است.

با توجه جدول ۱، عبارت‌های GO، به همه خوشه‌های دوبعدی منتسب‌نشده‌اند؛ باوجوداین، ممکن است این خوشه‌های دوبعدی ژن‌هایی را شامل شوند که با دانش فعلی، برای کارکردهای خاصی، ناشناخته مانده باشند. ممکن است با مطالعهٔ بیشتر دربارهٔ این خوشه‌های دوبعدی، به یافته‌های زیستی جدیدی دست‌یابیم.

شابالین و همکاران در سال ۲۰۰۹، روش LAS را در مقایسه با سایر روش‌های خوشه‌بندی، در مجموعه داده‌های بیان ژنی سرطان پستان و سرطان ریه، به‌کار-بردند. آنها نشان‌دادند که این روش در تشخیص زیرنوع‌های بیماری، از روش‌های خوشه‌بندی سلسله‌مراتبی و k-میانگین، بهتر عمل‌کرده‌است (۱۲).

اُکادا و همکاران در سال ۲۰۰۷، الگوریتم BiModule را برای خوشه‌بندی دوبعدی داده‌های بیان ژنی مخمرِ Saccharomyces cerevisiae *گسچ و همکاران*، به‌کار-بردند؛ آنها کارآیی این روش را (در یافتن خوشه‌های دوبعدی) در مقایسه با روش‌های خوشه‌بندی دوبعدی ISA، OPSM، Bimax، Samba، CC و xMotif بررسی-کردند و به‌منظور اعتبارسنجی زیستی خوشه‌های دوبعدی به‌دست‌آمده، از ابزار وِب FuncAssociate استفاده‌کردند. همهٔ خوشه‌های دوبعدی به‌دست‌آمده به روش BiModule، دست‌کم یک عبارت غنی‌شده معنی‌دار در سطح معنی‌داری ٪۰/۰۰۱ را دربرمی‌گرفت؛ روش‌های ISA، OPSM، Bimax، نیز درصد بالایی از خوشه‌های دوبعدی غنی‌شده را در سطوح معنی‌داری مختلف، نشان‌دادند (حدود ٪۹۰ تا ٪۱۰۰ در روش OPSM، حدود ٪۷۲ تا ٪۹۹ در روش Bimax و حدود ٪۸۰ تا ٪۹۱ در روش ISA) (۲۱).

فدل ال-اکوا و همکاران در سال ۲۰۰۹، بسته نرم‌افزاری (Automatic Gene Ontology; AGO) را برای اعتبارسنجی و مقایسه روش‌های خوشه‌بندی دوبعدی معرفی‌کردند. آنها به‌منظور بررسی این ابزار، روش‌های خوشه‌بندی دوبعدی ISA، OPSM، CC و BiVisu و روش خوشه‌بندی k-میانگین را در داده‌های بیان ژنی مخمرِ Saccharomyces cerevisiae *گسچ و همکاران* به‌کاربردند. در بررسی آنها، روش OPSM، درصد بالایی از خوشه‌های دوبعدی غنی‌شده کارکردی را در همه سطوح معنی‌داری نشان‌داد (از ٪۸۵ تا ٪۱۰۰)؛ البته این بیشتر به دلیل تعداد کم خوشه‌های دوبعدی تولیدشده (۲ خوشه) توسط این روش است؛ روش ISA نیز درصدی به‌نسبت بالا از خوشه‌های دوبعدی غنی‌شده را نشان-می‌دهد (۱۳).

تابع امتیاز LAS و رویکرد جستجو را می‌توان به آرایه‌هایی با ابعاد بالاتر، به‌طور نمونه، ماتریس‌های سه-بعدی داده‌ها، گسترش‌داد.

نتایج اعتبارسنجی زیستی الگوریتم LAS، در داده‌های بیان ژنی مخمر در پژوهش حاضر و همچنین نتایج حاصل از به‌کارگیری این روش، در داده‌های بیان ژنی سرطان پستان و سرطان ریه، اثربخشی معیار «میانگین-درایه‌های بزرگ» و رویکرد جستجوی LAS را تصدیق-می‌کند؛ البته، باید به این نکته توجه‌داشت که معیار «میانگین-درایه‌های بزرگ»، تنها یکی از معیارهایی است

که می‌توان در تحلیل اکتشافی داده‌های بیان ژنی به‌کار-برد. در مواردی، ممکن است که معیارها و روش‌های دیگر خوشه‌بندی دوبعدی، شناخت بیشتری درباره مجموعه داده‌های موردنظر، فراهم‌سازد؛ همچنین، ممکن است مواردی پیش‌آید که در آنها، استفاده از این معیار مناسب نباشد و روش‌های دیگر، اطلاعاتی ارزشمندتر را در اختیار محقق قراردهد. به دلیل اهمیت‌یافتن بهترین روش‌ها در تحلیل داده‌های بیان ژنی و لزوم تفسیر پزشکی و بالینی خوشه‌های حاصل، پیشنهادمی‌شود که تحقیقات بیشتری در قالب گروه‌های کاری با تخصص‌های بالینی، ژنتیک، آمارزیستی و بیوانفورماتیک انجام‌پذیرد.

تشکر و قدردانی

این مقاله، حاصل طرح تحقیقاتی دانشکدهٔ پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی است که بدین‌وسیله از هیئت رئیسهٔ محترم دانشکدهٔ پیراپزشکی برای تأمین بودجهٔ طرح پژوهشی، صمیمانه سپاسگزاری‌می‌شود.

منابع

1- Knudsen S. Guide to analysis of DNA microarray data. 2 ed: John Wiley and Sons 2004.

2- Francesco B, Adam P, Ivan P, Silvia S, Andrea S, Livia T, et al. GEMMA - A Grid environment for microarray management and analysis in bone marrow stem cells experiments. Elsevier Science Publishers B. V. 2007:382-90.

3- Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics. 2002;18 Suppl 1:S136-44.

4- Johnson RA, Wichern DW. Applied multivariate statistical data analysis: Prentice Hall: Upper Saddle River, NJ 2002.

5- Gu J, Liu JS. Bayesian biclustering of gene expression data. BMC Genomics. 2008;9 Suppl 1:S4.

6- Hartigan JA. Direct clustering of a data matrix. Journal of the american statistical association (JASA). 1972;67(337):123-9.

7- Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 2000:93-103.

8- Lazzeroni L, Owen A. Plaid models for gene expression data. Citeseer 2002:61-86.

9- Ben-Dor A, Chor B, Karp R, Yakhini Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. J Comput Biol. 2003;10(3-4):373-84.

10- Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. Published by the IEEE CS, NN, and EMB Societies & the ACM 2004:24-45.

11- Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006 May 1;22(9):1122-9.

12- Shabalin AA, Weigman VJ, Perou CM, Nobel AB. Finding large average submatrices in high dimensional data. 2009:985-1012.

13- Al-Akwaa FM, Kadah YM. An automatic gene ontology software tool for bicluster and cluster comparisons. IEEE 2009:1.7-63.

14- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell. 2000;11(12):4241-57.

15- Shenga Q, Lemmensa K, Marchalab K, De Moora B, Moreaua Y. Query-driven biclustering of microarray data by Gibbs sampling: Internal report 05-33, Department of Electrical Engineering (ESAT-SCD-SISTA), Katholieke Universiteit Leuven, Belgium; 2005.

16- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 2000 [cited; Available from: http://genome-www.stanford.edu/yeast_stress/

17- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature Publishing Group 2000:25-9.

18- Day-Richter J, Harris MA, Haendel M. OBO-Edit an ontology editor for biologists. Oxford Univ Press 2007:2198.

19- Carbon S, Ireland A, Mungall CJ, Shu SQ, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. Oxford Univ Press 2009:288.

20- Cheng KO, Law NF, Siu WC, Liew AWC. Biclusters Visualization and Detection Using Parallel Coordinate Plots. AIP Conference Proceedings. 2007;952(1):114-23.

21- Okada Y, Fujibuchi W, Horton P. A Biclustering Method for Gene Expression Module Discovery Using a Closed Itemset Enumeration Algorithm. IPSJ Digital Courier. 2007;3:183-92.

22- Tchagang AB, Gawronski A, Berube H, Phan S ,Famili F, Pan Y. GOAL: a software tool for assessing biological significance of genes groups. BMC Bioinformatics.11:229.

Daneshvar

Medicine

Scientific-Research Journal of Shahed University Seventeenth Year, No.93 June, July 2011

Received: 14/2/2011
Last revised: 9/5/2011
Accepted: 14/5/2011

Biclustering of DNA microarray gene expression data by Large Average Submatrices Method

Hamid Alavi Majd[1*], Shima Younespoor[1], Farid Zayeri[1], Mostafa Rezaei Tavirani[2]

1.Department of Biostatistics, Paramedical Sciences Faculty, Shahid Beheshti Medical Sciences University, Tehran, Iran.
2.Department of Basic Sciences, Paramedical Sciences Faculty, Shahid Beheshti Medical Science University, Tehran, Iran.

E-mail: alavimajd@gmail.com

Abstract

Background and Objective: In recent years, DNA microarray technology has become a central tool in genomic research. Using this technology, which made it possible to simultaneously analyze expression levels for thousands of genes under different conditions, massive amounts of information will be obtained. While traditional clustering methods, such as hierarchical and K-means clustering have been shown to be useful in analyzing microarray data, they have some limitations. These methods assume that a gene or an experimental condition can be assigned to only one cluster and a gene belongs to a group of genes that are coexpressed under all conditions. Therefore, to overcome these shortcomings, biclustering methods are used. The purpose of this paper was to evaluate the efficiency of a biclustering method in analyzing yeast gene expression data.

Materials and Methods: In this study, Large Average Submatrices (LAS) method has been used to analyze the yeast Saccharomyces cerevisiae expression dataset, provided by Gasch et al. (2000). The dataset contains 2993 genes and 173 different experimental conditions. In this study, the software packages such as LAS, JMP and GOAL has been used for analyzing data.

Results: Results showed that the LAS method is able to produce biologically and statistically relevant biclusters.

Conclusion: This study showed that LAS can be used to discover biologically significant subsets of genes under subsets of conditions for microarray data analysis.

Key words: Biclustering, Gene expression data, DNA microarray, Gene ontology, Large Average Submatrices (LAS)