هفتمین همایش ملی بیو تکنولوژی
جمهوری اسلامی ایران
زمان: ۲۱ تا ۲۳ شهریورماه ۱۳۹۰
مکان: تهران، پژوهشگاه نیرو
The 7th National Biotechnology Congress of I.R. Iran
12-14 Sep. 2011, Niroo Research Institute, Tehran, Iran

مقایسه روش‌های نرمال‌سازی و تعیین ژن در داده‌های میکروآرایه cDNA

سمانه فاضلی فارسانی[1*] - محمود اخوان مهدوی[2]
گروه مهندسی شیمی، دانشگاه فردوسی مشهد،
مشهد، میدان آزادی، پردیس دانشگاه فردوسی، دانشکده مهندسی، کد پستی ۹۱۷۷۹۴۸۹۴۴

چکیده
بیان ژن، پدیده‌ای اساسی در حیات موجودات زنده می‌باشد. از مهمترین روش‌های تعیین بیان ژن در سطح رونویسی، تکنولوژی میکروآرایه‌های cDNA بوده که قادرند بیان ده‌ها هزار ژن را بطور همزمان اندازه‌گیری کنند. اما آزمایش‌های میکروآرایه از ابتدا تا تولید داده‌های خام، در معرض تغییرات نامطلوبی بوده که ناشی از بایاس‌های آزمایشگاهی و تکنیکی هستند. جهت حذف یا حداقل کردن این تغییرات، روش‌های آماری تحت عنوان تکنیک‌های نرمال‌سازی با فرضیات آماری و بیولوژیکی متفاوت پیشنهاد شده است. پنج روش نرمال‌سازی به صورت منفرد و ترکیبی(در کل یازده روش) جهت پردازش مجموعه داده‌ی ApoAI بر اساس شناسایی هشت ژن دارای تنظیم پایین مقایسه شدند. نتایج آنالیزها نشان داد که روش نرمال‌سازی twoDloess به صورت منفرد و ترکیبی با روش median بهترین عملکرد را بر روی این مجموعه داده‌ها نشان می‌دهد. همچنین استفاده از P-value به تنهایی نتایج بهتری از کاربرد همزمان P-value و تغییرات چند برابری (Fold change) در مرحله‌ی تعیین ژن‌های با بیان متفاوت (Feature selection) حاصل می‌کند.

کلمات کلیدی: بیان ژن، میکروآرایه، نرمال‌سازی، تعیین ژن

مقدمه
نگاهی اجمالی به جریان اطلاعات سلول نشان می‌دهد که اطلاعات در مرحله‌ی رونویسی، از ژن‌ها (DNA) به RNA پیام‌رسان (mRNA) و سپس در مرحله‌ی ترجمه به پروتئین‌ها انتقال یافته که با مطالعه‌ی آنها می‌توان اطلاعات ژنتیکی فراوانی به دست آورد (۳). از معتبرترین روش‌ها در سطح رونویسی برای تعیین میزان بیان ژن‌ها، تکنولوژی میکروآرایه‌های cDNA بوده که با اندازه‌گیری همزمان بیان ده‌ها هزار ژن، خاموش (down-regulated) یا روشن (up-regulated) بودن آنها را در شرایط محیطی مشخص تعیین می‌سازند. در آزمایش میکروآرایه، بر روی آرایه‌هایی از جنس نایلون یا شیشه، نقاطی (اسپات) به طور منظم قرار گرفته‌اند که هر نقطه نماینده یک ژن است (پروب). پروب‌ها، الیگومرهای DNA بوده که با رشته‌های cDNA نشان‌دار واکنش داده و با استفاده از چاپگر بر روی سطح آرایه درون اسپات‌ها چاپ می‌شوند (۱). از آنجاییکه پروب‌های cDNA دو رشته‌ای بوده آرایه‌ی موردنظر برای تک‌رشته‌ای شدن پروب‌ها و عملیات هیبریداسیون، گرم یا تیمار قلیایی می‌گردد. سپس به منظور آماده‌سازی دو نمونه‌ی (تارگت) آزمایشی و کنترل، mRNA نمونه‌ها از طریق رونویسی معکوس به cDNA تبدیل می‌شوند. معمولاً نمونه‌ی آزمایشی در حین رونویسی معکوس با رنگ قرمز (Cy5) و نمونه‌ی کنترل با رنگ سبز (Cy3) نشان‌دار می‌گردند. سپس نمونه‌ها با یکدیگر ترکیب شده و بر روی سطح آرایه ریخته می‌شوند. دو نمونه به صورت رقابتی با پروب‌ها اتصال برقرار کرده و نمونه‌ای که بیان بالاتری برای یک پروب خاص دارد متصل می‌شود. بعد از شستشوی تارگت‌های هیبرید نشده، شدت رنگ‌ها با استفاده از اسکنر اندازه‌گیری می‌گردد. اسپات قرمز رنگ نشان می‌دهد که ژن موردنظر درون آن اسپات، در نمونه‌ی آزمایشی نسبت به نمونه‌ی کنترلی بیان بالاتر (رونوشت mRNA بیشتر) داشته و اسپات

[1]. دانشجوی کارشناسی ارشد، گروه مهندسی شیمی. samanefazeli@gmail.com
[2]. استادیار گروه مهندسی شیمی. mahdavi@ferdowsi.um.ac.ir

سبز رنگ، خلاف این را نشان می‌دهد. اسپات زرد رنگ بیانگر بیان یکسان ژن در هر دو نمونه است در حالیکه اسپات سیاه رنگ نشان می‌دهد که ژن در هیچکدام از نمونه‌ها بیان نشده است (۳).

بنابراین آزمایش میکروآرایه از ابتدا تا انتها در مراحلی نظیر استخراج mRNA، نشان‌دار کردن، اسکن و پردازش داده‌ها در معرض تغییرات ناخواسته‌ای بوده که ناشی از بایاس‌ها یا ناهماهنگی‌های آزمایشگاهی و تکنیکی می‌باشند. جهت حذف یا حداقل کردن برخی از این تغییرات، روش‌های آماری با عنوان نرمال‌سازی با فرضیات آماری و بیولوژیکی متفاوت استفاده می‌شوند که اکثر اوقات هدف، اصول و محدودیت‌های آن نادیده گرفته شده است. تفاوت در خواص فیزیکی رنگ‌ها، بازده‌های نشان‌دار کردن و اسکن دو رنگ با تنظیمات متفاوت اسکنر، باعث بایاس رنگ (dye bias) یا عدم تعادل بین رنگ‌ها می‌شوند (۶). در اکثر موارد بایاس رنگ ناشی از تفاوت در شدت‌های رنگ بوده به این معنا که بایاس رنگ اسپات تاریک متفاوت از اسپات روشن است. همچنین بایاس فضایی (spatial bias) ناشی از ناهمگونی بین پرینت‌تیپ‌های مورد استفاده در فرایند ساخت آرایه، تغییرات در سطح آرایه و شستشوی غیر یکنواخت است (۸). قرار دادن پروب‌ها بر روی آرایه با استفاده از پین‌ها انجام می‌گیرد. تعداد و آرایش پین‌ها بر روی هد چاپگر، ساختار اصلی ردیف‌ها و ستون‌های موجود در هر پرینت‌تیپ را تشکیل می‌دهند (۳). بایاس مقیاس (scale bias) نیز از تفاوت در مواد آزمایشگاهی، تجهیزات به ویژه تغییر در تنظیمات لوله فتومولتی‌پلایر در اسکنر و شرایط محیطی در زمان پردازش آرایه ایجاد می‌گردد (۶و۹).

بعد از نرمال‌سازی داده‌ها، آنالیزهای بیشتری جهت به دست آوردن نتایج معنادار بیولوژیکی نیاز است. در حقیقت، هدف اصلی آزمایش میکروآرایه، تعیین ژن‌هایی با بیان متفاوت (DEG) در شرایط مختلف بیولوژیکی و یا بالینی می‌باشد. به منظور مقایسه‌ی روش‌های نرمال‌سازی و تعیین ژن، تأثیر روش‌ها بر شناسایی ژن‌هایی با بیان متفاوت در مجموعه داده ApoAI بررسی گردید. این پژوهش در راستای نیاز مبرم به اطلاعات گسترده در روش‌های پردازش داده‌های میکروآرایه به خصوص در ایران با توجه به ورود تکنولوژی میکروآرایه‌ها و استفاده روزافزون آن انجام گرفته است.

مواد و روش‌ها

مجموعه داده ApoAI: این داده شامل ۱۶ آرایه با ۶۳۸۴ ژن می‌باشد. هشت آرایه حاوی نمونه‌ی آزمایشی مربوط به موشی که ژن ApoAI آن خاموش شده و هشت آرایه دیگر مربوط به نمونه‌ی موش کنترلی است. در طراحی آزمایش از طراحی غیرمستقیم استفاده شده که هر دو نمونه‌ی نشان‌دار شده با رنگ قرمز با نمونه‌ی رفرنس نشان‌دار شده با رنگ سبز هیبرید می‌شوند. هدف از آزمایش، شناسایی ژن‌های با بیان متفاوت در نمونه آزمایشی نسبت به کنترل است. بررسی‌های انجام گرفته روی این داده نشان می‌دهد که هشت ژن از ۶۳۸۴ ژن موجود در آرایه در موشی با ژن خاموش شده ApoAI در مقایسه با موش‌های نرمال، دارای تنظیم پایین هستند (۲).

روش نرمال‌سازی میانه: در روش‌های کلی فرض می‌شود که شدت‌های سبز و قرمز با ضریب ثابتی به یکدیگر مرتبط می‌شوند یعنی $R=k\times G$. بنابراین جهت نرمال‌کردن داده‌ها مطابق با رابطه (۱)، پارامتر c از نسبت‌های لگاریتمی کم شده و مرکز توزیع نسبت‌های لگاریتمی (میانه یا میانگین) تصحیح می‌شود. استراتژی‌های مختلفی برای تعیین پارامتر c وجود دارد. در روشی که توسط زین و همکاران (۲۰۰۱) ارائه گردید با فرض بیان متفاوت تعداد کمی از ژن‌ها در دو نمونه، ثابت نرمال‌سازی به عنوان یک پارامتر تمایل مرکزی (نظیر میانه) از نسبت‌های لگاریتمی روی هر آرایه در نظر گرفته شد (۱۱).

$$[\log_2(R_i/G_i)]_{shifted} = \log_2(R_i/G_i) - c = \log_2(R_i/G_i) - \log_2(k) = \log_2 R/(kG) \qquad (1)$$

روش نرمال‌سازی LOESS: این روش به سه نوع، LOESS کلی(Gloess)، پرینت‌تیپ LOESS (PTloess) و LOESS دوبعدی (twoDloess) تقسیم می‌گردد. روش وزن‌دهی LOESS، تابعی برای برازش استوار منحنی‌های هموار در رگرسیون موضعی است. فرض اعمال شده در روش LOESS نیز عبارت است از اینکه اکثر ژن‌های روی آرایه تحت شرایط مورد بررسی به صورت متفاوت بیان نمی‌شوند (۶). بنابراین میزان شدت رنگ‌های سبز و قرمز برای اکثر ژن‌ها یکسان بوده و نسبت لگاریتمی آنها ($M=\log_2(R/G)$) صفر خواهد شد. در ذیل رابطه مربوط به نرمال‌سازی Gloess آورده شده است.

$$[\log_2(R_i/G_i)]_{shifted} = \log_2(R_i/G_i) - c(A) = \log_2(R_i/G_i) - \log_2(k(A)) = \log_2 R_i/[k(A)G] \quad (۲)$$

c(A)، پارامتر برازش LOESS روی کل داده‌های موجود در نمودار پراکندگی است. دقت این روش به کسری از داده‌ها که میزان شدت ($A= 1/2\log_2(R\times G)$) آنها به شدت اسپات مورد بررسی نزدیکتر هستند بستگی داشته که برای هموارسازی در هر نقطه استفاده می‌شوند. در نرمال‌سازی PTloess پارامتر برازش از طریق داده‌های هر پرینت‌تیپ تخمین زده می‌شود. رابطه (۳) چگونگی نرمال‌سازی با روش PTloess را نشان می‌دهد که $c_p(A)$، پارامتر برازش LOESS برای نمودار پراکندگی در p امین پرینت‌تیپ می‌باشد.

$$[\log_2(R_i/G_i)]_{shifted} = \log_2(R_i/G_i) - c_P(A) = \log_2(R_i/G_i) - \log_2(k_P(A)) = \log_2 R/[k_p(A)G] \quad (۳)$$

در نرمال‌سازی twoDloess از تابع LOESS دوبعدی استفاده شده که می‌تواند با برازش این منحنی به داده‌ها، وابستگی نسبت‌های لگاریتمی به موقعیت‌های فیزیکی اسپات‌ها را برآورد کند(۶).

$$N = M - LOESS(r,c) \quad (۴)$$

$LOESS(r,c)$، منحنی LOESS برازش شده است که تابعی از موقعیت ردیف (r) و ستون (c) اسپات‌های روی آرایه می‌باشد. این روش، تغییرات فضایی را با یک سطح هموار دو بعدی به جای استفاده از یک منحنی مدل می‌کند(۶).

روش نرمال‌سازی مقیاس: روش نرمال‌سازی مقیاس(PTMAD) می‌تواند برای تصحیح مقیاس داده‌ها بین پرینت‌تیپ‌های مختلف انجام شود(۱۰). در این روش نیز فرض عدم بیان متفاوت اکثر ژن‌ها اعمال می‌گردد، بنابراین باید تغییرات (واریانس یا مقیاس) در داده‌های مختلف از پرینت‌تیپ‌ها، یکسان باشد. همچنین فرض می‌شود که $Ms=M/s$ نسبت‌های شدت نرمال‌شده با این روش بوده که M نشان دهنده‌ی نسبت‌های لگاریتمی شدت ژن‌ها قبل از نرمال‌سازی و s نیز فاکتور مقیاس بوده و از طریق پارامتر انحراف مطلق از میانه (Median Absolute Deviation) محاسبه می‌گردد. می‌توان برآورد مناسبی از پارامتر مقیاس را در i امین گروه پرینت‌تیپ در رابطه (۵) مشاهده نمود.

$$\hat{s}_i = (MAD_i)/\left[\sqrt[I]{\Pi_{i=1}^{I} MAD_i}\right] \quad (۵)$$

انحراف مطلق از میانه نیز به صورت ذیل در رابطه (۶) تعریف می‌گردد.

$$MAD_i = median\left\{\left|M_{ij} - median_j(M_{ij})\right|\right\} \quad (۶)$$

تعیین ژن‌هایی با بیان متفاوت: ساده‌ترین روش جهت مشخص کردن تغییرات بیان ژن یا عدم تغییر و میزان آن در نمونه‌ی آزمایشی نسبت به کنترل، تقسیم میزان بیان هر ژن در نمونه‌ی آزمایشی به میزان بیان همان ژن در نمونه‌ی کنترل است. این روش تغییرات چندبرابری (Fold change)، از تکنیک‌های قدیمی در تعیین DEG ها بوده که کماکان کاربرد وسیعی در رتبه‌بندی ژن‌ها دارد. اگر این نسبت از یک آستانه‌ی تعیین‌شده بیشتر باشد، ژن به صورت متفاوت بیان شده است. معمولاً از آستانه‌ی دو برابری در تنظیم بالا و پایین ژن به عنوان یک حد پایه (Cut off) استفاده می‌شود. از آزمون t نیز می‌توان استفاده کرد که در این پژوهش یک آماره‌ی t تصحیح شده (آماره t بیز تجربی) به کار رفته و P-valueهای تصحیح شده برای مقایسات چندگانه نیز برای نشان دادن معناداری تغییرات بیان ژن در مجموعه داده‌ها تعیین می‌گردند(۷). آنالیزهای انجام شده در این پژوهش در نرم‌افزار R انجام گرفته و از بسته‌های نرم‌افزاری Limma و marray موجود در پایگاه اطلاعاتی Bioconductor (http://www.bioconductor.org) برای نرمال‌سازی و تعیین ژن‌هایی با بیان متفاوت استفاده گردید (۶و۱۰).

نتایج و بحث

نرمال‌سازی، فرایندی مهم در پردازش داده‌های میکروآرایه جهت به دست آوردن نتایجی با معناداری بیولوژیکی بالا است. حذف انواع بایاس‌های موجود در داده از طریق یک روش نرمال‌سازی امکانپذیر نیست. بنابراین باید از توالی از روش‌ها برای حداقل کردن تغییرات نامطلوب موجود استفاده کرد. در این پژوهش برخی از روش‌ها به تنهایی و به صورت ترکیبی، از طریق تعیین تعداد ژن‌های با بیان متفاوت در آزمایش ApoAI مقایسه گردید. روش‌های نرمال‌سازی به صورت ترکیبی قادر به حذف

همزمان بایاس‌های رنگ و فضایی و یا بایاس‌های رنگ و مقیاس می‌باشند. روش PTloess همانطور که Wu و همکاران (۲۰۰۵) اشاره کرده‌اند، به عنوان روش حذف بایاس محلی رنگ استفاده گردید (۹). بعد از نرمال‌سازی داده‌ها، ژن‌های با بیان متفاوت از طریق دو روش تعیین گردید که نتایج در جدول ۱ گزارش شده است. نتایج این جدول نشان می‌دهد که در این داده‌ها استفاده از P-value به تنهایی برای شناسایی ژن‌های موردنظر بهتر از ترکیب P-value با تغییرات چندبرابری عمل کرده و تعداد بیشتری از ژن‌های موردنظر شناسایی می‌شود. استفاده از $P<0.01$ نیز به جز در روش PTMAD به صورت منفرد و ترکیب با twoDloess نتایج مشابهی با $P<0.05$ داد (نتایج آورده نشده است). هر چند که پترسون و همکاران (۲۰۰۶) نشان دادند که ترکیب پارامتر معناداری آماری ($P<0.01$ یا $P<0.05$) با تغییرات چند برابری (۱/۵، ۲ یا ۴) بهتر از P-value به تنهایی عمل می‌کند (۵). می‌توان چنین نتیجه گرفت که پیشنهاد روشی خاص برای تعیین ژن‌هایی با بیان متفاوت برای داده‌های مختلف میکروآرایه امکانپذیر نیست. بلکه به دلیل دخیل بودن طیف وسیعی از پارامترها در این نوع آزمایش‌ها بهتر است همراه با آزمایش میکروآرایه، بیان برخی از ژن‌ها را با RT-PCR یا تکنیک‌های دیگر اندازه گرفت تا معیار مناسبی جهت انتخاب بهترین روش در داده‌ی موردنظر داشت.

جدول ۱-تعداد ژن‌های با بیان متفاوت بعد از روش‌های مختلف نرمال‌سازی و تعیین ژن

روش‌های نرمال‌سازی	None	Median	Gloess	PTloess	twoDloess	PTMAD	median+twoDloess	Gloess+twoDloess	PTloess+twoDloess	median+PTMAD	Gloess+PTMAD	PTloess+PTMAD
P-value<0.05	۸	۸	۸	۸	۸	۷	۸	۸	۸	۸	۸	۸
Pvalue<0.05+fold change(2>and<-2)	۳	۷	۵	۵	۸	۶	۸	۵	۴	۷	۶	۵

کل روش‌های نرمال‌سازی بجز PTMAD هشت ژن موردنظر را با $P<0.05$ به درستی شناسایی کردند. البته عدم نرمال‌سازی (none) و ترکیب‌های Gloess+twoDloess و median+PTMAD با وجود شناسایی هشت ژن موردنظر، چند ژن دیگر را نیز به اشتباه به عنوان DEG معرفی نمودند. در هنگام استفاده از $P<0.05$ همراه با تغییرات چند برابری، روش twoDloess در بین روش‌های منفرد و median+twoDloess در بین روش‌های ترکیبی، بیشترین تعداد درست DEG را شناسایی کرد. روش median برای تصحیح نوعی از بایاس رنگ پیشنهاد شده که البته قادر به حذف بایاس‌های وابسته به شدت و فضایی نبوده و به عنوان یک روش پایه استفاده می‌گردد. twoDloess نیز برای بایاس‌های فضایی پیشنهاد شد. قابل ذکر است که تنها پژوهش انجام گرفته بر روی نرمال‌سازی داده‌های میکروآرایه در ایران توسط نقی‌زاده و همکاران (۲۰۰۵) انجام شده که تنها با محاسبه‌ی میانگین و انحراف استاندارد شدت‌های بیان بعد از روش محلی LOESS و مقیاس نتیجه گرفتند که داده‌ها بعد از نرمال‌سازی از پایداری بیشتری برخوردار خواهند بود (۴).

مراجع

1. **Barrett, J. C., Kawasaki, E.S., 2003**. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. DDT. 3: 134-141.
2. **Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., Rubin, E. M., 2000**. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. Genome Research. 10: 2022-2029.
3. **Karakach, T. K., Flight, R. M., Douglas, S., 2010.** An introduction to DNA microarrays for gene expression analysis. Chemometrics and Intelligent Laboratory Systems. 1: 28–52.
4. **Naghizadeh, M. M., Hajizadeh, E., Kazemnejad, A., 2005.** cDNA microarray data normalization. Iranian journal of biotechnology. 1: 55-63.
5. **Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T. M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., Walker, S.J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J. C., Tong, W., Shi, L., Wolfinger, R. D., 2006**. Performance comparison of one-color and two-colorplatformswithin theMicroArrayQuality Control (MAQC) project. Nat. Biotechnol. 9: 1140–1150.
6. **Smyth, G. K., Speed, T., 2003**. Normalization of cDNA microarray data. Methods. 4: 265-273
7. **Smyth, G. K., 2004**. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology. 1: Article 3
8. **Uchida, S., Nishida, Y., Satou, K., Muta, S., Tashiro, K., Kuhara, S., 2005**. Detection and normalization of biases present in spotted cDNA microarray data: a composite method addressing dye, intensity-dependent, spatially-dependent, and print-order biases. DNA Research, 12:1-7.
9. **Wu, W., Xing, E. P., Myers, C., Mian, S., Bissell, M. J., 2005**. Evaluation of normalization methods for cDNA microarray data by k-NN classification. BMC Bioinformatics. 191:
10. **Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., Speed, T. P., 2002**. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research. 4: e15.
11. **Zien, A., Aigner, T., Zimmer, R., Lengauer, T., 2001**. Centralization: a new method the normalization of gene expression data. Bioninformatics. 1: 323-331.

Comparison of Normalization and Feature Selection Methods on cDNA Microarray Data

Gene expression is a fundamental process in life span and microarray technology is the most important technique in measuring expression of tens of thousands of genes at transcription level. Microarray experiments are subject to undesirable variations throughout the steps for generating raw data. The variations are caused by technical and experimental inconsistencies or biases. In order to remove or reduce the variations, normalization techniques with different statistical or biological assumptions are applied. In this study, five normalization methods were implemented separately or in combination with each other (eleven methods in total) for processing ApoAI dataset. Comparison was performed based on eight genes that are known a priori to be down-regulated. Results indicated twoDloess individually and in combination with median show the best performance in normalization of the dataset. Also P-value alone gives much better performance than P-value with fold-change in feature selection step.