شناسایی ژن‌های مرتبط با بقا در سرطان کلیه با استفاده از روش مؤلفه‌های اصلی لاسو

مریم السادات دهقانی: دانشجوی کارشناسی ارشد آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی تهران، تهران، ایران. dehghani_mar@yahoo.com
* **محمود رضا گوهری**: دانشیار، گروه آمارزیستی، دانشکده بهداشت، دانشگاه علوم پزشکی ایران، تهران، ایران (*نویسنده مسئول). gohar_ma@yahoo.com
سهیلا خداکریم: استادیار، گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران. lkhodakarim@gmail.com

تاریخ دریافت: ۹۴/۳/۲۶ تاریخ پذیرش: ۹۴/۷/۷

چکیده

زمینه و هدف: یافتن ژن‌های مرتبط با بقا بر مبنای داده‌های بیان ژن، یک کاربرد مهم داده‌های ریزآرایه می‌باشد. هدف از مطالعه حاضر شناسایی ژن‌های مرتبط با بقای بیماران مبتلا به سرطان متعارف سلول‌های کلیوی با استفاده از داده‌های بیان ژن حاصل از ریزآرایه است.

روش کار: این مطالعه از نوع تحلیل بقای داده‌های ابعاد بالا است. ۱۷۷ نمونه بیمار مبتلا به سرطان متعارف سلول‌های کلیوی (conventional Renal Cell Carcinoma: cRCC) و برای هر فرد ۱۴۸۱۴ ژن مورد بررسی قرارگرفته است. برای تشخیص ژن‌های مرتبط با بقا از روش مؤلفه‌های اصلی لاسو (Lassoed Principal Components: LPC) استفاده شده است که از اطلاعات بیان همه ژن‌ها برای محاسبه امتیاز یک ژن استفاده می‌نماید. در نهایت برای تعیین تعداد ژن‌های معنادار از معیار نرخ تشخیص کاذب (False Discovery Rate: FDR) استفاده شده است. تجزیه و تحلیل آماری با استفاده از نرم‌افزار R انجام شده است.

یافته‌ها: بر اساس یافته‌های این مطالعه استفاده از نقطه برش ۰/۰۰۱ برای معیار FDR و بررسی ۱۰۴۱ ژن مرتبط با وضعیت بقای بیماران cRCC که به ترتیب اهمیت قدر مطلق امتیازهای LPC بالاتری دارند، امکان بروز کمترین خطا را فراهم آورده است.

نتیجه‌گیری: در این پژوهش پس از رتبه‌بندی ژن‌ها توسط امتیاز LPC برحسب میزان تغییرات بیان مرتبط با وضعیت بقای بیماران cRCC، ۱۱ ژن از مهم‌ترین ژن‌های مرتبط با بقا شناسایی شدند. امتیازهای LPC این ۱۱ ژن منفی هستند بنابراین با افزایش بیان این ژن‌ها بقای بیماران cRCC افزایش یافته است و به عبارت دیگر افزایش بیان این ژن‌ها فاکتورهای محافظتی این بیماران به شمار می‌آیند.

کلیدواژه‌ها: داده ابعاد بالا، ریزآرایه، بیان ژن، تحلیل بقا

مقدمه

در دهه اخیر پیشرفت و گسترش فن‌آوری‌های جدید در حوزه پزشکی تغییر شگرفی در داده‌های زیست پزشکی ایجاد نموده است. به‌عنوان نمونه، روش‌های تعیین الگوهای ژنتیکی سبب ایجاد و تولید داده‌های بسیار زیاد برای هر فرد شده است. این داده‌ها درک ما از فرآیندهای بیولوژیکی و بیماری‌هایی مانند سرطان را به‌کلی تغییر داده‌اند؛ این پیشرفت‌ها سبب ارزشمند شدن داده‌هایی شده است که در آن‌ها تعداد ویژگی‌ها بیشتر از تعداد مشاهدات است که این داده‌ها با عنوان ابعاد بالا شناخته می‌شوند. داده‌های حاصل از فناوری ریزآرایه که ژن‌ها (بیان ژن‌ها) ویژگی‌ها و بیماران مشاهدات آن هستند؛ مجموعه داده با متغیرهای ابعاد بالا است. هنگامی‌که برای همین بیماران نتایج برآمد بقا هم موجود باشد آنگاه می‌توان ژن‌های مرتبط با بقای بیماران را شناسایی نمود. ویژگی‌هایی که مرتبط با بقا تشخیص داده می‌شوند، سپس در آزمایش‌هایی که به‌منظور بهتر فهمیدن فرآیند بیولوژیکی که منجر به برآمد نهایی بیماری می‌شود، مورد بررسی قرار می‌گیرند؛ همچنین این ویژگی‌ها می‌توانند برای پیش‌بینی بقا یا طبقه‌بندی بقای بیماران جدید استفاده شوند (۱). ارائه اهدافی برای پیشرفت‌های دارویی جدید نیز می‌تواند از دیگر مزایای یافتن ژن‌های مرتبط با بقا باشند (۲).

رویکرد استاندارد برای مدل بندی داده‌های بقا هنگامی‌که تعداد مشاهدات بیشتر از تعداد ویژگی‌ها است، برازش مدل خطرات متناسب کاکس بر داده مورد نظر است (۳، ۴). در حالی که برازش این مدل بر داده‌های با متغیرهای ابعاد بالا مناسب نیست و حتی هنگامی‌که تعداد مشاهدات

[Downloaded from rjms.iums.ac.ir on 2025-02-16]

و تعداد متغیرها به هم نزدیک هستند این مدل توان پایینی دارد. بنابراین روش‌هایی جهت کاهش بعد داده‌ها به‌منظور وفق دادن مدل خطرات متناسب کاکس با داده‌های با متغیرهای ابعاد بالا به کار می‌روند (۵).

در این مطالعه یک مدل رگرسیونی برای داده‌های بقایی که در آن‌ها تعداد متغیرهای توضیحی با ابعاد بالا است، جهت یافتن ژن‌های مرتبط با بقا معرفی می‌شود. این روش برای اولین بار توسط ویتن و تیب شیرانی در سال ۲۰۰۸ ارائه گردید و بهبود قابل‌توجهی در روش‌های معمول تحلیل بقا فراهم آورد. در این مدل به‌منظور محاسبه امتیاز یک ژن از اطلاعات بیان همه ژن‌ها استفاده می‌شود. از دیدگاه علوم زیستی، ژنی که به‌درستی با یک برآمد (مانند بقا) مرتبط است جزئی از یک فرآیند زیستی است که شامل تعداد بسیار زیادی از ژن‌هاست که این ژن‌ها به دلیل هم تنظیمی، الگوی بیان مشابهی نشان می‌دهند. ازآنجاکه ژن‌های هم تنظیم بیان ژن‌های همبسته دارند، بنابراین اگر بیان یک ژن با بیان گروه بزرگی از ژن‌ها که مرتبط با زمان بقای بیماران تشخیص داده شده‌اند، همبستگی داشته باشد، شانس بیشتری وجود دارد که آن ژن مرتبط با زمان بقا تشخیص داده شود (۱). هدف اصلی از این مقاله رتبه‌بندی ژن‌های مورد بررسی برحسب شدت ارتباط با وضعیت بقای بیماران است به‌گونه‌ای که با افزایش قدر مطلق امتیاز ژن‌ها، ارتباط آن‌ها با وضعیت بقای بیماران افزایش یابد. در این راستا از روش LPC (Lassoed Principal Components) استفاده شد و با برازش این مدل رگرسیونی بر داده‌های بیماران cRCC (conventional Renal Cell Carcinoma)، به هر ژن امتیازی تعلق گرفت.

روش کار

رایج‌ترین روش برای شناسایی ویژگی‌های مرتبط با زمان بقا استفاده از امتیازهای تک متغیره کاکس است که در معادله (۱) بیان شده است. به‌منظور محاسبه این امتیاز برای هر ویژگی x_j مدل خطرات متناسب کاکس تک متغیره برازش می‌شود، آماره امتیاز یا امتیاز کاکس حاصل از مدل، معیاری کمّی از میزان پیش‌بینی بقا توسط آن متغیر ارائه می‌نماید(۳, ۵).

هنگامی‌که $\ell(\beta)$ لگاریتم درستنمایی نسبی مدل رگرسیونی کاکس تک متغیره باشد، آماره امتیاز بدین صورت محاسبه می‌گردد(۳, ۵, ۶):

(۱)

$$s_j = \left(\frac{dl(0)}{d\beta_j}\right) \Big/ \left(\frac{d^2 l(0)}{d\beta_j^2}\right)^{\frac{1}{2}} = \frac{\sum_{r\in D}\left(x_j^r - \frac{1}{n_r}\sum_{i\in R_r} x_j^i\right)}{\left[\sum_{r\in D}\frac{1}{n_r}\sum_{i\in R_r}\left(x_j^i - \frac{1}{n_r}\sum_{k\in R_r} x_j^k\right)^2\right]^{\frac{1}{2}}}$$

r اندیس زمانی است که پیشامد رخ داده است (t_r) و n_r تعداد افراد در معرض خطر در زمان t_r هستند. همچنین x_j^r متغیر توضیحی j ام از مشاهده‌ای است که در زمان t_r پیشامد برایش رخ داده است و R_r مجموعه مشاهدات در معرض خطر در زمان t_r هستند.

مقدار زیاد قدر مطلق s_j نشان‌دهنده ارتباط ویژگی x_j با برآمد بقا است و علامت امتیاز کاکس نشان می‌دهد که ژن مورد نظر با افزایش بقا ارتباط دارد یا کاهش آن. هنگامی‌که امتیاز کاکس منفی است، بیان ژن بالاتر نشان‌دهنده بقای طولانی‌تر است درحالی‌که امتیاز کاکس مثبت نشان می‌دهد، بیان ژن بالاتر نشان‌دهنده بقای کمتر است (۷-۹).

به‌منظور ارزیابی دقیق‌تر معناداری یک ژن، روش‌های زیادی هستند که از اطلاعات ژن‌های دیگر استفاده می‌نمایند؛ روش LPC یکی از این روش‌هاست که برای تشخیص ژن‌های مرتبط با بقای بیماران از تجزیه مقادیر منفرد ماتریس داده‌های بیان ژن استفاده می‌نماید. طریقه محاسبه امتیاز LPC با استفاده از امتیاز کاکس بدین ترتیب است:

اگر T برداری از امتیازهای کاکس برای ژن‌های مورد بررسی باشد و $v_1,...,v_n \in \mathbb{R}^p$ بردارهای منفرد از راست برای ماتریس داده‌های بیان ژن X باشند، امتیاز LPC توسط معادله $\hat{T} = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_i v_i$ به دست می‌آید، در حالی که

[Downloaded from rjms.iums.ac.ir on 2025-02-16]

(۲)

$$\hat{\beta} = \arg\min_{\beta} \left\{ \left\| T - \beta_0 - \sum_{i=1}^{n} v_i \beta_i \right\|^2 + \lambda \sum_{i=1}^{n} |\beta_i| \right\}.$$

$\hat{T}$ مقادیر برازش یافته‌ای است که از رگرسیون امتیازهای کاکس روی آرایه‌های ویژه (بردار ویژه‌های مشاهدات) ماتریس داده‌های بیان ژن، با اعمال تاوان L_1 (لاسو) حاصل شده است(۵). پارامتر تنظیم‌کننده $\lambda \geq 0$ موجود در تاوان L_1 توسط روش اعتبارسنجی به‌گونه‌ای برآورد می‌گردد که سبب کاهش خطای امتیازهای حاصل شود.

ویژگی‌هایی که قدر مطلق امتیاز آن‌ها بیشتر باشد، معنادارتر در نظر گرفته می‌شوند و ویژگی‌های بالای لیست معناداری، مرتبط با بقا تشخیص داده می‌شوند (۱). به‌منظور یافتن ویژگی‌های مرتبط با بقا، به دنبال لیستی از ویژگی‌های معنادار هستیم که شامل تعداد کمی مثبت کاذب باشد. برای رسیدن به این منظور از معیار FDR (False Discovery Rate) استفاده می‌گردد. معیار FDR عبارت است از نسبت مورد انتظار ویژگی‌های مثبت کاذب از میان ویژگی‌هایی که معنادار شناسایی شده‌اند (۱۰-۱۵).

در این مطالعه از مجموعه داده ژائو و همکاران استفاده شده است که شامل ۱۴۸۱۴ بیان ژن (p) استخراج شده از بافت سرطانی برای ۱۷۷ نمونه (n) بیمار مبتلا به سرطان متعارف سلول‌های کلیوی است که بین سال‌های ۱۹۸۵ تا ۲۰۰۳ در بیمارستان دانشگاه آمیا سوئد مورد جراحی خارج نمودن کامل کلیه قرار گرفته بودند، هم‌چنین در این مجموعه داده زمان بقا و وضعیت سانسور برای هر یک از بیماران نیز موجود است. داده‌های کامل حاصل از ریزآرایه که در این مجموعه داده آمده است در پایگاه داده GEO (Gene Expression Omnibus) با آدرس اینترنتی *ww.ncbi.nlm.nih.*gov و با شماره دسترسی *GSE3538* موجود است. داده‌های بیان ژن، تبدیل لگاریتمی سطوح بیان ژن هستند و در قالب یک ماتریس n × p مورد بررسی قرار می‌گیرند که سطرهای آن نماینده بیماران و ستون‌های آن نشانگر ژن‌ها هستند. هم‌چنین به‌منظور نرمال‌سازی داده‌های بیان ژن برای خارج نمودن خطاهای سیستماتیک، ستون‌های این ماتریس مرکزی شده‌اند و دارای میانگین صفر هستند. تجزیه و تحلیل آماری با استفاده از نرم‌افزار R انجام شده است.

یافته‌ها

برای تعیین پارامتر تنظیم شده بهینه برای تاوان لاسو از روش اعتبارسنجی ابتدا مجموعه داده مورد نظر به دو مجموعه داده‌های آزمایشی و آزمون تقسیم شد به‌گونه‌ای که هریک از مجموعه داده‌های آزمایشی و آزمون ۵۰ درصد مجموعه داده کامل را تشکیل می‌دادند. تقسیم داده‌ها به مجموعه‌های آموزشی و آزمون به‌صورت تصادفی ۱۰ مرتبه تکرار شد و پارامتر تنظیم شده بهینه برای تاوان لاسو $\hat{\lambda} = ۱۸/۴۶۴۸$ برآورد شده است. استفاده از این برآورد بر ای پارامتر انقباضی منجر به غیر صفر شدن ضریب رگرسیونی هفت بردار ویژه به کار گرفته شده در محاسبه امتیازهای LPC گردیده است. برآورد ضرایب همراه با بردار

جدول ۱- برآورد ضرایب غیر صفر بردار ویژه‌های به کار رفته در محاسبه امتیاز LPC

اندیس بردار ویژه (i)	مقدار ضریب رگرسیونی برآورده شده ($\hat{\beta}_i$)
۱	۲۰/۵۶۵
۲	۴۶/۴۵۳
۳	۲۱/۰۳۸
۴	-۵۷/۰۰۹
۶	-۱۲/۲۵۸
۱۴	۶/۷۴۶
۱۵	۳/۸۰۹

[Downloaded from rjms.iums.ac.ir on 2025-02-16]

جدول ۲- نقطه برش های FDR و تعداد ژن های با FDR کمتر از نقطه برش

نقطه برش FDR	تعداد ژن های با LPCFDR < نقطه برش
۰/۰۰۱	۱۰۴۱
۰/۰۱	۱۱۰۴
۰/۰۵	۱۴۹۵
۰/۱	۲۰۶۶
۰/۲	۳۵۹۲
۰/۴	۷۴۲۹

FDR=LPCFDR امتیازهایLPC

جدول ۳- امتیازهای LPC، FDR و نام ژن های مرتبط با بقا که امتیاز بزرگتر از ۴ دارند

ردیف	امتیاز LPC	LPCFDR	نام ژن
۱	-۶ / ۱۰۶	۰	TMEM27
۲	-۵ / ۳۳۵	۰	ACSM2B
۳	-۴ / ۹۵۱	۰	NPR3
۴	-۴ / ۷۹۳	۰	ACSM2A
۵	-۴ / ۳۸۰	۰	GSTA2
۶	-۴ / ۲۶۷	۰	RGS5
۷	-۴ / ۱۷۸	۰	ENPP2
۸	-۴ / ۱۰۳	۰	APOM
۹	-۴/۰۶۲	۰	SLC5A12
۱۰	-۴/۰۵۴	۰	PDK4
۱۱	-۴/۰۳۲	۰	ADAMTS9

ویژه‌های مربوط به آن در جدول ۱ آمده است.

با توجه به نتایج جدول ۱، دومین و چهارمین آرایه ویژه حاصل از ماتریس داده‌ها، بیشترین قدر مطلق ضرایب رگرسیونی را در مدل رگرسیونی لاسو دارا می‌باشند. با استفاده از ضرایب رگرسیونی جدول ۱ امکان محاسبه امتیاز LPC هر یک از ژن‌ها فراهم شد. پس از محاسبه امتیاز LPC ژن‌ها، برای یافتن تعداد ژن‌های معنادار از نقطه برش‌های متفاوت برای معیار FDR استفاده شد و برای هر یک از آن‌ها تعداد ژن‌های مرتبط با وضعیت بقا یافت شد که نتایج در جدول ۲ گزارش شده است.

بر اساس نتایج جدول ۲ با افزایش مقدار معیار FDR، تعداد ژن‌های معنادار و در نتیجه امکان بروز خطا در ژن‌های معنادار افزایش یافته بود. با در نظر گرفتن نقطه برش ۰/۰۰۱، تعداد ۱۰۴۱ ژن مرتبط با وضعیت بقا شناسایی شدند که از این میان تقریباً یک ژن به اشتباه تشخیص داده شده بود، در حالی که با نقطه برش ۰/۰۱ برای معیار FDR، از میان ۱۱۰۴ ژن مرتبط با وضعیت بقا شناسایی شده تقریباً ۱۱ ژن غیرمرتبط با وضعیت بقا در میان این ژن‌ها وجود دارد؛ بنابراین انتخاب نقطه برش ۰/۰۰۱ برای معیار FDR و بررسی ۱۰۴۱ ژن مرتبط با وضعیت بقا امکان بروز کمترین خطا را فراهم آورده است.

در جدول ۳ امتیاز LPC و نام ژن‌هایی که قدر مطلق امتیاز LPC بیشتر از ۴ دارند همراه با FDR امتیازهای LPC آن‌ها گزارش شده است.

بحث و نتیجه گیری

با توجه به تحلیل انجام شده در مطالعه حاضر که با استفاده از روش LPC به شناسایی ژن های مرتبط با بقای بیماران cRCC پرداخته شد، ژن ها برحسب میزان تغییرات بیان مرتبط با وضعیت بقای بیماران رتبه بندی شدند. ژن هایی که امتیاز LPC بزرگتر از ۴ دارند و نام آنها در جدول ۳ ذکر شده است از مهمترین ژن های مرتبط با بقای بیماران cRCC شناسایی شده اند و بقیه ژن هایی

[Downloaded from rjms.iums.ac.ir on 2025-02-16]

که تغییرات بیان آنها در این مطالعه بررسی شده است در درجه اهمیت کمتری قرار دارند و در رتبه‌های بعدی از لحاظ تاثیر بر بقا قرار می گیرند. این ۱۱ ژن همگی امتیاز LPC منفی دارند که حاکی از این مطلب است که با افزایش بیان این ژن‌ها بقای بیماران cRCC افزایش یافته است و به عبارت دیگر افزایش بیان این ژن ها فاکتورهای محافظتی این بیماران به شمارمی آیند. ازآنجاکه اکثر ژن های مورد بررسی در این مطالعه امتیاز LPC منفی کسب نموده اند، بنابراین در بیماران با بقای طولانی‌تر، افزایش بیان داشته‌اند. جالب توجه است که همین الگوی تغییرات بیان ژن در نتایج حاصل از مقاله تاکاهاشی و همکاران و همچنین مقاله ژائو و همکاران نیز وجود دارد (۱۶، ۱۷). در مقاله تاکاهاشی و همکاران ۵۱ ژن از عوامل پیش‌بینی کننده بقای بیماران cRCC معرفی شده بودند که افزایش بیان بخش قابل‌توجهی از این ژن‌ها همراه با افزایش بقای این بیماران بوده است (۱۶). در مقاله ژائو و همکاران نیز در ۲۵۹ ژنی که تغییرات بیان آن‌ها پیش‌بینی کننده بقای بیماران cRCC معرفی شده بود، همین الگوی تغییرات بیان دیده می‌شد. هم‌چنین در مقایسه نتایج حاصل از این پژوهش با مقاله ژائو و همکاران مشاهده می‌شود که از میان ۱۱ ژن ذکر شده در جدول ۳ همگی به جز ژن APOM که در رتبه هشتم این ژن‌ها قرار دارد، در میان ۲۵۹ ژنی هستند که در مقاله ژائو و همکاران از عوامل پیش‌بینی کننده بقای بیماران cRCC شناسایی شده بودند. همچنین این ۱۱ ژن با ۴۵ ژنی که در مقاله وسلی و همکاران با استفاده از امتیاز کاکس حاصل از مدل خطرات متناسب کاکس مرتبط با بقای بیماران cRCC شناسایی شده بودند، هم‌پوشانی نداشتند که احتمالاً به این دلیل است که مجموعه داده مورد بررسی در مطالعه وسلی و همکاران شامل گروهی از بیماران با تومورهای متعارف (Conventional) و غیرمتعارف (Nonconventional) از لحاظ بافت شناسی بود (۱۸). درحالی‌که تومورهایی که در مطالعه ما بررسی شده‌اند تنها شامل تومورهای متعارف هستند.

ازآنجاکه تحقیقات آزمایشگاهی که به‌منظور بررسی ژن‌های مرتبط با یک فنوتایپ طراحی می‌شوند اهداف متفاوتی دارند، متناسب با هدف هر تحقیقی که در ارتباط با بررسی ژن‌ها در بقای بیماران مبتلا به cRCC باشد، محقق می‌تواند ژن‌ها را از بالای لیست معناداری حاصل شده از این مطالعه با ترتیب اختصاص داده شده به آن‌ها به‌عنوان ژن‌های هدف برای بررسی‌های آزمایشگاهی استفاده نماید. به‌طورکلی با توجه به یافته‌های این مطالعه می‌توان نتیجه گرفت که تغییرات بیان برخی از ژن‌هایی که مهم‌ترین آن‌ها در جدول ۳ ذکرشده‌اند، بقای بیماران cRCC را تحت تأثیر قرار می‌دهد و حتی در مواردی که این بیماری از لحاظ کلینیکی پنهان باشد، بررسی بیان ژن‌های ذکرشده می‌تواند در تشخیص و یا یافتن درمان مناسب این بیماران بسیار راهگشا باشد.

از نقاط قوت این پژوهش این است که امکان استفاده از الگوهای ژنتیکی بیماران cRCC را جهت تشخیص و درمان آن‌ها فراهم می‌نماید درحالی‌که به‌طور معمول جهت ارزیابی بقای این بیماران از متغیرهای کلینیکی درجه تومور (Grade)، مرحله تومور(Stage) و وضعیت کارایی (Performance status) (سلامت عمومی) بیمار، استفاده می‌شود(۱۷). قابل ذکر است که اقبال کم مراکز تحقیقاتی ژنتیک کشور به فناوری ریزآرایه و عدم وجود مجموعه داده‌های حاصل از ریزآرایه برای جامعه بیماران ایرانی از عمده مشکلات پیشرو در این پژوهش بوده است که امکان استفاده از این فناوری پرکاربرد و جدید را در حوزه سلامت کشور با محدودیت مواجه نموده است.

تقدیر و تشکر

این تحقیق از پایان‌نامه کارشناسی ارشد رشته آمار زیستی دانشکده بهداشت دانشگاه علوم پزشکی تهران استخراج شده است که بدین‌وسیله از حمایت‌های مالی دانشگاه علوم پزشکی تهران تشکر و قدردانی می‌گردد.

[Downloaded from rjms.iums.ac.ir on 2025-02-16]

منابع

1. Witten DM, Tibshirani R. Testing significance of features by lassoed principal components. The annals of applied statistics. 2008;2(3):986.
2. Hastie TJ, Tibshirani RJ, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: Springer; 2009.
3. Klein J, Moeschberger M. Survival analysis: statistical methods for censored and truncated data. Springer-Verlag, New York, NY. 2003.
4. Kleinbaum D, Klein M. Survival Analysis: A self-learning text, 2005. New York, Springer-Verlag; 2011.
5. Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. Statistical methods in medical research. 2010;19(1):29-51.
6. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data: John Wiley & Sons; 2011.
7. Mandruzzato S, Callegaro A, Turcatel G, Francescato S, Montesco MC, Chiarion-Sileni V, et al. A gene expression signature associated with survival in metastatic melanoma. J Transl Med. 2006;4:50.
8. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001;98(19):10869-74.
9. Chu G, Li J, Narasimhan B, Tibshirani R, Tusher V. Significance Analysis of Microarrays Users Guide and Technical Document. 2001.
10. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.
11. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4(4):210.
12. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Statistical Science. 2003:71-103.
13. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19(3):368-75.
14. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64(3):479-98.
15. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003;100(16):9440-5.
16. Takahashi M, Rhodes DR, Furge KA, Kanayama H-o, Kagawa S, Haab BB, et al. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. Proceedings of the National Academy of Sciences. 2001;98(17):9754-9.
17. Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD. Gene expression profiling predicts survival in conventional renal cell carcinoma. PLoS medicine. 2005;3(1):e13.
18. Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, et al. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. Proceedings of the National Academy of Sciences. 2003;100(12):6958-63.

[Downloaded from rjms.iums.ac.ir on 2025-02-16]

Original Article http://rjms.iums.ac.ir

Identification of related genes with survival in renal carcinoma by using lassoed principal components method

Maryam Dehghani, MSc of Biostatistics, Tehran University of Medical Sciences, Tehran, Iran. dehghani_mar@yahoo.com
*__Mahmood Reza Gohari__, Associate professor of Biostatistics, Iran University of Medical Sciences, Tehran, Iran (*Corresponding author). gohar_ma@yahoo.com
Soheila Khodakarim, Assistant Professor of Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran. lkhodakarim@gmail.com

Abstract

Background: Identification of correlated genes with survival by gene expression data is an important application of microarray data. The purpose of this study is to identify correlated genes with survival of conventional renal cell carcinoma (cRCC) patients based on gene expression profiles.

Methods: This study is a survival analysis with high dimensional covariates and containing 14814 gene expression measurements from 177 patients with cRCC. Lassoed principal components (LPC) method is used for identification associated genes with survival. LPC score uses information of all of gene expressions for computation a gene score. Finally False Discovery Rate (FDR) method is used to identify significant genes. Statistical analysis is done with using the R software.

Results: The lowest error is satisfied with using the cutoff 0.001 for FDR criteria and with studying 1041 related genes with survival of cRCC patients.

Conclusion: 11 genes are identified as most significant genes with survival of cRCC patients, after ranking the genes with their LPC scores with regard to their differentially expressions. The LPC scores of these 11 genes are negative, so increase of these gene expressions are related to increase of the survival of cRCC patients and in the other words the increase of these gene expressions are protective factors in cRCC patients.

Keywords: High dimensional data, Microarray, Gene expression, Survival analysis

[Downloaded from rjms.iums.ac.ir on 2025-02-16]