دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

مقدمه‌ای بر بیوانفورماتیک

پاییز ۱۴۰۳

استاد: علی شریفی زارچی
مسئول تمرین: آراد ملکی

مهلت ارسال بدون تاخیر: ۵ دی | **پاسخ‌نامه تمرین سوم** | مهلت ارسال نهایی: ۸ دی

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روزهای مشخص‌شده است.
- در طول ترم، برای هر تمرین می‌توانید تا ۴ روز تأخیر داشته باشید و در مجموع حداکثر ۸ روز تأخیر مجاز خواهید داشت. توجه داشته باشید که تأخیر در تمرین‌های عملی و تئوری به صورت مشترک محاسبه می‌شود. پس از اتمام تاخیرهای مجاز، می‌توانید با تاخیری ساعتی ۱ درصد تمرین خود را ارسال کنید.
- حتماً تمرین‌ها را بر اساس موارد ذکرشده در صورت سوالات حل کنید. در صورت وجود هرگونه ابهام، آن را در صفحه تمرین در سایت کوئرا مطرح کنید و به پاسخ‌هایی که از سوی دستیار آموزشی مربوطه ارائه می‌شود، توجه کنید.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- فایل پاسخ‌های سوالات نظری را در قالب یک فایل *pdf* به فرمت $HW۳\_[STD\_ID].pdf$ آپلود کنید.
- گردآورندگان تمرین: آراد ملکی، علی حاجی‌صادقیان، آرزو پاک‌سرشت، سینا نمازی

سوالات نظری (۱۰۰ نمره)

۱. (۱۰ نمره) به سوالات زیر در رابطه با Transcription پاسخ دهید:

الف) تفاوت‌های اساسی بین فرآیند رونویسی در پروکاریوت‌ها و یوکاریوت‌ها چیست و این تفاوت‌ها چگونه بر تنظیم ژن و پیچیدگی آن تأثیر می‌گذارند؟

ب) چگونه اپی ژنتیک، از جمله متیلاسیون DNA و تغییرات هیستونی، می‌تواند بر الگوی رونویسی ژن‌ها تأثیر بگذارد و در نتیجه در تنظیم بیان ژن دخیل باشد؟

پ) توضیح دهید که چگونه فاکتورهای رونویسی اختصاصی می توانند بر شروع رونویسی در سلول های یوکاریوتی تأثیر بگذارند و نقش آن ها در تنظیم بیان ژن چیست؟

ت) چگونه مهارکننده‌های رونویسی می‌توانند به عنوان داروهای ضد سرطان عمل کنند؟ یک مثال از چنین مهارکننده‌ای را توضیح دهید و مکانیزم عمل آن را بیان کنید.

ث) مفهوم Alternative Splicing چیست و چگونه می‌تواند به تنوع پروتئینی منجر شود؟

حل.

الف) در پروکاریوت ها رونویسی و ترجمه هم زمان در سیتوپلاسم انجام می شود و معمولاً mRNA بدون تغییرات پس رونویسی (نظیر اسپلیسینگ) مستقیماً استفاده می گردد. در مقابل، یوکاریوت ها دارای هسته اند؛ رونویسی در هسته انجام و سپس mRNA با اضافه شدن کلاهک ۵′، دنباله پلی آ و حذف اینترون ها پردازش می شود. این فرایندهای اضافی و حضور عناصر تنظیمی پیچیده (نظیر پروموترهای پیشرفته و اینهنسرها) باعث افزایش دقت، تنظیم چندسطحی و تنوع بیان ژن در یوکاریوت ها می شود.

ب) متیلاسیون DNA و تغییرات شیمیایی در هیستون ها ساختار کروماتین را تغییر داده و دسترسی فاکتورهای رونویسی را به ژن ها تنظیم می کنند. در نتیجه با روشن یا خاموش کردن ژن ها، بیان ژنی را کنترل می کنند.

پ) فاکتورهای رونویسی اختصاصی در سلول های یوکاریوتی به توالی های خاصی در DNA متصل می شوند (مانند پروموترها و Enhancer ها) و با جذب یا دفع اجزای رونویسی، شروع رونویسی را تنظیم می کنند. آن ها نقش کلیدی در تنظیم بیان ژن دارند و پاسخ سلول به محرک های محیطی و سیگنال های درون سلولی را هماهنگ می کنند.

ت) مهارکننده های رونویسی با مسدود کردن فعالیت پروتئین های کلیدی درگیر در رونویسی ژن های مرتبط با رشد و تکثیر سلول های سرطانی، از پیشرفت سرطان جلوگیری می کنند. یک مثال از این داروها اکتینومایسین D است که با اتصال به DNA و جلوگیری از حرکت RNA پلی مراز، مانع از رونویسی mRNA می شود. این مکانیسم منجر به مهار رشد سلول های سرطانی می گردد.

ث) فرآیندی در RNA پردازش است که طی آن اگزون ها به روش های مختلف به هم متصل می شوند یا اینترون ها حذف می گردند. این فرایند امکان تولید چندین نوع mRNA از یک ژن را فراهم می کند، که منجر به ساخت پروتئین های مختلف از یک توالی ژنی واحد می شود. به این ترتیب، تنوع پروتئینی بدون افزایش تعداد ژن ها حاصل می شود.

۲. (۱۰ نمره) به سؤالات زیر در مورد تحلیل داده‌های ریزآرایه (Microarray) پاسخ دهید:

الف) تحلیل داده‌های ریزآرایه چیست و چه کاربردی در بیوانفورماتیک دارد؟ توضیح دهید چگونه این فناوری می‌تواند بیان ژن‌ها را اندازه‌گیری کند.

ب) اصول اساسی طراحی یک آزمایش ریزآرایه را توضیح دهید و بیان کنید چرا انتخاب نمونه‌ها اهمیت دارد.

پ) اصول عملکرد پروب‌ها در ریزآرایه چیست و چرا انتخاب آن‌ها در طراحی آزمایش مهم است؟

ت) مراحل کلی تحلیل داده‌های ریزآرایه چیست؟ برای هر مرحله توضیح مختصری ارائه دهید.

ث) تفاوت اصلی بین آرایه‌های یک رنگ (Single-Color Array) و دو رنگ (Dual-Color Array) چیست و چه مزایا یا معایبی دارند؟

ج) مزایای استفاده از ریزآرایه در مقایسه با روش‌های سنتی تحلیل بیان ژن چیست؟

چ) چه نوع اطلاعاتی از داده‌های ریزآرایه قابل استخراج است و چگونه این اطلاعات به درک عملکرد ژن‌ها کمک می‌کند؟

ح) یکی از مشکلات معمول در تحلیل داده‌های ریزآرایه، نویزهای پس‌زمینه است. چه روش‌هایی برای کاهش این نویزها وجود دارد؟

خ) در تحلیل داده‌های ریزآرایه، چرا نرمال‌سازی ضروری است؟ دو روش نرمال‌سازی را توضیح دهید.

د) چگونه از داده‌های ریزآرایه برای شناسایی ژن‌های دیفرانسیلی بیان‌شده (Differentially Expressed Genes) استفاده می‌شود؟ مراحل این فرآیند را شرح دهید.

ذ) محدودیت‌های ریزآرایه در مقایسه با تکنیک‌های جدیدتر مانند RNA-Seq چیست؟

حل.

الف) ریزآرایه یک فناوری است که به کمک آن می‌توان بیان همزمان هزاران ژن را اندازه‌گیری کرد. ریزآرایه‌ها مجموعه‌ای از پروب‌های DNA هستند که به یک سطح جامد متصل شده‌اند. نمونه‌های RNA یا DNA تک‌رشته‌ای به پروب‌ها هیبرید می‌شوند و سطح بیان ژن‌ها با اندازه‌گیری سیگنال‌های فلورسانس تعیین می‌شود. این تکنیک برای شناسایی سطح بیان ژن‌ها در شرایط مختلف استفاده می‌شود و و این امکان را می‌دهد که بتوان ژن‌های دیفرانسیلی بیان‌شده (DEGs) را در شرایط بیمار و سالم شناسایی کرد.

ب) مراحل طراحی شامل انتخاب نمونه‌ها، استخراج RNA با کیفیت بالا، برچسب‌گذاری فلورسانس، هیبریداسیون و تحلیل سیگنال‌های فلورسانس است. انتخاب دقیق نمونه‌ها برای اطمینان از نتایج قابل اعتماد و نماینده از شرایط زیستی بسیار مهم است.

پ) پروب‌ها قطعات کوتاهی از DNA هستند که به توالی مکمل RNA یا DNA هدف متصل می‌شوند. انتخاب پروب‌های اختصاصی برای جلوگیری از اتصال غیراختصاصی و بهبود دقت آزمایش حیاتی است. معمولاً الیگونوکلئوتیدها به دلیل قابلیت طراحی دقیق برای هر ژن خاص و ثبات بالا به‌عنوان پروب استفاده می‌شوند. در واقع، هنگام هیبریداسیون، فقط RNA یا cDNA مکمل به پروب‌ها متصل می‌شود و سیگنال فلورسانس تولید می‌کند.

ت) ۱. جمع‌آوری داده‌های خام: ثبت سیگنال‌های فلورسانس. ۲. نرمال‌سازی داده‌ها: حذف نویزها و قابل‌مقایسه کردن نمونه‌ها. ۳. شناسایی ژن‌های دیفرانسیلی بیان‌شده: تحلیل آماری برای یافتن ژن‌های با تغییر معنادار. ۴. تحلیل مسیرها: بررسی مسیرهای زیستی فعال. ۵. تفسیر نتایج: مقایسه با داده‌های زیستی موجود.

ث) عملکرد کلی آرایه‌های یک رنگ و دو رنگ مشابه است. در ریزآرایه‌های دو رنگ، دو نمونه بیولوژیکی (نمونه آزمایشی و نمونه کنترل) با رنگ‌های مختلف فلورسنت، معمولاً Cy3 (سیانین ۳) و Cy5 (سیانین ۵)، برچسب‌گذاری می‌شوند.

پس از هیبریداسیون رقابتی، اندازه‌گیری فلورسانس به طور جداگانه برای هر رنگ انجام شده و نشان‌دهنده فراوانی هر ژن در یک نمونه (نمونه آزمایشی، Cy5) نسبت به نمونه دیگر (نمونه شاهد، Cy3) است. داده‌های هیبریداسیون به صورت نسبت سیگنال‌های فلورسنت Cy5/Cy3 در هر پروب گزارش می‌شود.

در این روش، مزیت اصلی امکان مقایسه مستقیم بین دو نمونه روی یک اسلاید است؛ اما عیب آن احتمال تداخل رنگ و نیاز به دقت بالا در انجام آزمایش است.

در مقابل، در ریزآرایه‌های یک رنگ، هر نمونه به طور جداگانه برچسب‌گذاری شده و در یک ریزآرایه مستقل هیبرید می‌شود. این روش مقدار مطلق فلورسانس را برای هر پروب به دست می‌دهد. به طور دقیق‌تر، فقط یک نمونه روی هر اسلاید قرار می‌گیرد که باعث وضوح بیشتر داده‌ها و حذف تداخل رنگ می‌شود.

ج) امکان تحلیل همزمان هزاران ژن، صرفه‌جویی در زمان و هزینه، و دقت بالا در شناسایی بیان ژن‌های با سطح پایین از مزایای ریزآرایه است.

چ) داده‌های ریزآرایه سطح بیان ژن، ژن‌های دیفرانسیلی بیان‌شده، و شبکه‌های تنظیمی را نشان می‌دهند. این اطلاعات برای درک عملکرد ژن‌ها و مسیرهای زیستی مفید است.

ح) روش‌هایی مانند نرمال‌سازی پس‌زمینه (RMA)، حذف پروب‌های ضعیف و فیلترگذاری داده‌ها به کاهش نویز کمک می‌کند.

خ) نرمال‌سازی داده‌ها برای کاهش نویز، تنظیم مقیاس داده‌ها، و مقایسه‌پذیری بین نمونه‌ها ضروری است. دو روش نرمال‌سازی:

۱. درون‌آرایه‌ای: تنظیم تفاوت رنگ‌های فلورسانس (مانند Lowess).

۲. بین‌آرایه‌ای: یکسان‌سازی توزیع سیگنال‌ها در آرایه‌ها (مانند Quantile Normalization).

د) فرآیند شامل شناسایی ژن‌های دیفرانسیلی بیان‌شده از طریق تحلیل آماری (مانند t-test یا ANOVA)، تعیین ژن‌هایی با adjusted p-value پایین‌تر از یک آستانه مشخص، و تهیه لیستی از ژن‌های هدف برای تحلیل بیشتر است. ژن‌هایی که سطح بیان آن‌ها در شرایط مختلف تفاوت معنی‌دار دارند شناسایی می‌شوند. این ژن‌ها می‌توانند به‌عنوان ژن‌های دیفرانسیلی بیان‌شده معرفی شوند.

ذ) محدودیت‌های ریزآرایه شامل عدم حساسیت به توالی‌های ناشناخته، دقت پایین در اندازه‌گیری ژن‌های با بیان کم، و نیاز به طراحی پروب‌ها است. مقایسه مزایای این روش شامل هزینه کمتر، توان عملیاتی بالا و مناسب بودن برای ژنوم‌های با اطلاعات کامل است.

۳. (۲۰ نمره) در این تمرین به بررسی اثر دیابت نوع ۲ بر سطح بیان یک ژن خاص می‌پردازیم. ژن GLUT۴ نقش کلیدی در انتقال گلوکز به داخل سلول‌ها و ارتباط مستقیم با حساسیت به انسولین دارد. کاهش بیان این ژن ممکن است یکی از عوامل مهم در ایجاد یا پیشرفت دیابت نوع ۲ باشد. می‌خواهیم بدانیم آیا سطح بیان ژن

GLUT۴ در بیماران مبتلا به دیابت نوع ۲ نسبت به افراد سالم تغییر کرده است یا خیر. برای این منظور، سطح بیان ژن GLUT۴ در دو گروه از افراد اندازه‌گیری شده است. گروه اول شامل نمونه‌های سالم و گروه دوم شامل نمونه‌های بیمار مبتلا به دیابت نوع ۲ است. سطح بیان این ژن در این نمونه‌ها پس از نرمال‌سازی به شرح زیر گزارش شده است:

۱۲/۳	۱۲/۲	۱۱/۷	۱۲/۰	۱۲/۴	۱۱/۹	۱۲/۱	۱۲/۵	۱۱/۸	۱۲/۳

جدول ۱: سطح بیان ژن GLUT۴ در نمونه‌های سالم

۱۱/۰	۱۰/۷	۱۰/۸	۱۱/۱	۱۰/۹	۱۰/۶	۱۰/۷	۱۱/۰	۱۰/۸	۱۰/۵

جدول ۲: سطح بیان ژن GLUT۴ در نمونه‌های بیمار

الف) مشخص کنید که کدام آماره برای بررسی تفاوت بین دو گروه استفاده می‌شود و چرا مناسب است؟

ب) فرض صفر (H_0) و فرض مقابل (H_1) در این آزمایش را تعریف کنید.

ج) آزمون t را انجام دهید. آماره t را محاسبه کنید و مقدار p-value را به دست آورید.

د) آیا مقدار p-value کمتر از سطح معنی‌داری (۰/۰۵) است؟ توضیح دهید که آیا فرض صفر رد می‌شود یا خیر.

ه) براساس نتایج به دست آمده، توضیح دهید که آیا دیابت نوع ۲ تأثیری بر سطح بیان ژن GLUT۴ داشته است؟

حل.

الف) آماره t مناسب است. چون دو گروه مستقل هستند، متغیر کمی است، هدف مقایسه میانگین‌هاست و با فرض نرمال بودن داده‌ها.

ب)
- فرض صفر: میانگین سطح بیان ژن GLUT۴ در دو گروه (سالم و بیمار) یکسان است.
- فرض مقابل: میانگین سطح بیان ژن GLUT۴ در دو گروه (سالم و بیمار) متفاوت است.

ج)

$$\bar{X} = \frac{\sum x_i}{n},$$
$$s^2 = \frac{\sum (\bar{X} - x_i)^2}{n-1},$$
$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$
$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

$$\bar{X}_1 = \frac{11 + 10/7 + 10/8 + 11/1 + 10/9 + 10/6 + 10/7 + 11 + 10/8 + 10/5}{10} = 10/81,$$
$$\bar{X}_2 = \frac{12/3 + 12/2 + 11/7 + 12 + 12/4 + 11/9 + 12/1 + 12/5 + 11/8 + 12/3}{10} = 12/12.$$

$$s_1^2 = \frac{(۱۰/۸۱ - ۱۱)^2 + (۱۰/۸۱ - ۱۰/۷)^2 + \cdots + (۱۰/۸۱ - ۱۰/۵)^2}{۹}$$
$$= \frac{۰/۰۳۶۱ + ۰/۰۱۲۱ + ۰/۰۰۰۱ + ۰/۰۸۴۱ + ۰/۰۰۸۱ + ۰/۰۴۴۱}{۹}$$
$$+ \frac{۰/۰۱۲۱ + ۰/۰۳۶۱ + ۰/۰۰۰۱ + ۰/۰۹۶۱}{۹} = ۰/۰۳۶۶.$$

$$s_2^2 = \frac{(۱۲/۱۲ - ۱۲/۳)^2 + (۱۲/۱۲ - ۱۲/۲)^2 + \cdots + (۱۲/۱۲ - ۱۲/۳)^2}{۹}$$
$$= \frac{۰/۰۳۲۴ + ۰/۰۰۶۴ + ۰/۱۷۶۴ + ۰/۰۱۴۴ + ۰/۰۷۸۴ + ۰/۰۴۸۴}{۹}$$
$$+ \frac{۰/۰۰۰۴ + ۰/۱۴۴۴ + ۰/۱۰۲۴ + ۰/۰۳۲۴}{۹} = ۰/۰۷۰۷.$$

$$t = \frac{۱۲/۱۲ - ۱۰/۸۱}{\sqrt{\frac{۰/۰۳۶۶}{۱۰} + \frac{۰/۰۷۰۷}{۱۰}}}$$
$$= \frac{۱/۳۱}{۰/۱۰۳۶} = ۱۲/۶۵.$$

$$df = \frac{\left(\frac{۰/۰۳۶۶}{۱۰} + \frac{۰/۰۷۰۷}{۱۰}\right)^2}{\frac{\left(\frac{۰/۰۳۶۶}{۱۰}\right)^2}{۹} + \frac{\left(\frac{۰/۰۷۰۷}{۱۰}\right)^2}{۹}}$$
$$= \frac{۰/۰۰۰۱۱۵۱۳۲۹}{\frac{۰/۰۰۰۰۱۳۳۹۵۶}{۹} + \frac{۰/۰۰۰۰۴۹۹۸۴۹}{۹}} = ۱۶/۳۵.$$

برای $t = ۱۲/۶۵$ و $df = ۱۶/۳۵$:

$$p = ۲ \cdot (۱ - CDF(|t|, df)).$$

مقدار p محاسبه شده برابر است با:

$$p = ۷/۲۹ \times ۱۰^{-۱۰}.$$

د) مقدار p-value محاسبه شده از سطح معناداری ۰/۰۵ خیلی کمتر است. در نتیجه فرض صفر رد می‌شود.

ه) بله، با توجه به اینکه فرض صفر رد شد، پس میانگین سطح بیان ژن 4GLUT در دو گروه (سالم و بیمار) متفاوت است.

۴. (۲۰ نمره) در این سوال به بررسی ارتباط بین دو دارو و خطر حمله قلبی می‌پردازیم. یک مطالعه بالینی طراحی شده است تا ارتباط بین مصرف دو داروی Aspirin و Ibuprofen و کاهش خطر ابتلا به حمله قلبی (Myocardial Infarction) بررسی شود. این مطالعه شامل گروهی از افراد است که به طور تصادفی به دو دسته تقسیم شده‌اند: گروه اول از داروی Aspirin و گروه دوم از داروی Ibuprofen استفاده کرده‌اند. در پایان دوره‌ی مطالعه، محققان نتایج زیر را ثبت کردند:

گروه	داروی Aspirin	داروی Ibuprofen	مجموع
دچار حمله قلبی نشده‌اند	۲۲۰	۱۸۰	۴۰۰
دچار حمله قلبی شده‌اند	۸۰	۱۲۰	۲۰۰
مجموع	۳۰۰	۳۰۰	۶۰۰

فرضیات مطالعه:

H_0: هیچ ارتباطی بین مصرف دارو و خطر ابتلا به حمله قلبی وجود ندارد. به عبارت دیگر، شانس وقوع حمله قلبی برای دو دارو یکسان است.

H_1: شانس وقوع حمله قلبی برای افرادی که داروی Aspirin مصرف کرده‌اند کمتر از افرادی است که داروی Ibuprofen مصرف کرده‌اند.

الف) نسبت شانس (Odds Ratio) را محاسبه کنید. با استفاده از داده‌های جدول، نسبت شانس برای وقوع حمله قلبی در گروه‌های مصرف‌کننده دو دارو را محاسبه کنید. توضیح دهید که این نسبت چه مفهومی دارد.

ب) نتیجه محاسبات خود را در چارچوب فرضیه‌های صفر و جایگزین تفسیر کنید. آیا نسبت شانس به نفع داروی خاصی است؟

ج) از آزمون دقیق فیشر استفاده کنید تا p-value را برای این داده‌ها محاسبه کنید. توضیح دهید که این آزمون چگونه کار می‌کند و چرا در این مطالعه به کار می‌رود.

د) نتیجه آزمون را در چارچوب فرضیه‌ها تحلیل کنید. آیا مصرف داروی Aspirin تأثیر معناداری در کاهش خطر حمله قلبی دارد؟

حل.

الف)

$$\text{Ratio Odds} = \frac{\text{۱ group in event of Odds}}{\text{۲ group in event of Odds}}$$

$$\text{(Aspirin) Odds} = \frac{\text{(Aspirin) occurred attack Heart}}{\text{(Aspirin) occur not did attack Heart}} = \frac{۸۰}{۲۲۰} = ۰/۳۶۴$$

$$\text{(Ibuprofen) Odds} = \frac{\text{(Ibuprofen) occurred attack Heart}}{\text{(Ibuprofen) occur not did attack Heart}} = \frac{۱۲۰}{۱۸۰} = ۰/۶۶۷$$

$$\text{OR} = \frac{\text{(Aspirin) Odds}}{\text{(Ibuprofen) Odds}} = ۰/۵۴۶$$

نسبت شانس (Ratio Odds یا OR) یک معیار آماری است که برای مقایسه شانس وقوع یک رخداد در دو گروه مختلف استفاده می‌شود. این مفهوم به ویژه در مطالعات پزشکی و اپیدمیولوژی برای بررسی ارتباط میان یک عامل (مثل مصرف دارو) و یک پیامد (مثل بروز بیماری) به کار می‌رود.

تفسیر نسبت شانس:

- OR = ۱: شانس وقوع رخداد در دو گروه یکسان است (هیچ ارتباطی وجود ندارد).
- OR < ۱: شانس وقوع رخداد در گروه اول بیشتر از گروه دوم است.
- OR > ۱: شانس وقوع رخداد در گروه اول کمتر از گروه دوم است.

ب)

مصرف داروی خاص (Aspirin) تأثیری بر کاهش یا افزایش خطر وقوع حمله قلبی ندارد. ($OR = ۱$) : $H_۰$

مصرف داروی Aspirin تأثیر دارد ($OR \neq ۱$) : $H_۱$

نتیجه محاسبات: مقدار OR محاسبه‌شده برابر با ۰/۵۴۶ است. این مقدار کمتر از ۱ است، به این معنا که احتمال وقوع حمله قلبی با مصرف داروی Aspirin نسبت به Ibuprofen کاهش می‌یابد. پس مصرف Aspirin به نسبت Ibuprofen به نفع کاهش خطر حمله قلبی است.

ج) آزمون دقیق فیشر احتمال مشاهده جدول توافقی فعلی (یا جدولی با نسبت‌های شدیدتر) را در صورت درست بودن فرضیه صفر محاسبه می‌کند. این آزمون می‌تواند به طور دقیق بررسی کند که آیا رابطه معناداری بین دو متغیر (مثلاً استفاده از دارو و بروز نتیجه خاص) وجود دارد یا خیر.

د) مقدار p-value به دست آمده ۰/۰۰۰۷۱۲ است که از ۰/۰۵ کمتر است. یعنی فرض صفر رد می‌شود. پس می‌توان گفت که Ibuprofen نسبت به Aspirin در کاهش شانس حمله قلبی مؤثرتر است.

۵. (۲۰ نمره) به سؤالات زیر در مورد RNA-Seq و پایگاه‌های مرتبط پاسخ دهید:

الف) تکنیک RNA-Seq چیست و چگونه به مطالعه بیان ژن کمک می‌کند؟ مزایای این روش نسبت به تکنیک‌های سنتی مانند Microarray چیست؟

ب) تفاوت میان داده‌های raw و processed در RNA-Seq چیست؟ چرا هر دو نوع داده برای تحلیل‌ها اهمیت دارند؟

پ) پایگاه داده SRA چیست و چه نوع داده‌هایی در آن ذخیره می‌شود؟ چگونه می‌توان به داده‌های RNA-Seq در این پایگاه دسترسی پیدا کرد؟

ت) یک مطالعه مرتبط با سرطان در پایگاه SRA پیدا کنید که داده‌های RNA-Seq را ارائه داده باشد. Accession Number آن را مشخص کنید و مراحل جستجوی خود را توضیح دهید.

ث) با استفاده از ابزار NCBI SRA Toolkit داده‌های خام مطالعه‌ای که در بخش قبل یافتید را دانلود کنید. دستورات مورد استفاده برای این کار را بیان کنید.

ج) پایگاه داده Ensembl را بازدید کنید و یک ژن خاص مرتبط با بیماری (مثلاً BRCA۱ برای سرطان پستان) را پیدا کنید. داده‌های بیان ژن مرتبط با این ژن را جستجو و توضیح دهید چه اطلاعاتی ارائه می‌شود.

حل.

الف) تکنیک RNA-Seq روشی پیشرفته برای تحلیل بیان ژن است که با استفاده از توالی‌یابی نسل جدید (Next-Generation Sequencing) انجام می‌شود. در این روش، RNA استخراج‌شده از نمونه‌ها به cDNA تبدیل و سپس توالی‌یابی می‌شود تا میزان بیان ژن‌ها با دقت بسیار بالا مشخص شود. مزیت RNA-Seq نسبت به تکنیک‌های سنتی مانند Microarray در این است که RNA-Seq نیازی به پروب‌های از پیش طراحی‌شده ندارد، امکان شناسایی ترانسکریپت‌های جدید، انواع مختلف RNA و بیان ژن‌ها در مقیاس وسیع را فراهم می‌کند و حساسیت بیشتری در اندازه‌گیری بیان ژن دارد.

ب) داده‌های raw در RNA-Seq شامل اطلاعات خام توالی‌یابی‌شده هستند که معمولاً در قالب فایل‌هایی مانند FASTQ ارائه می‌شوند و مستقیماً از دستگاه توالی‌یاب به‌دست می‌آیند. داده‌های processed نتایج تحلیل‌های پردازشی روی داده‌های خام هستند که شامل اطلاعاتی مانند مقادیر نرمال‌شده بیان ژن و ترانسکریپت‌ها می‌شود. داده‌های خام برای بازتحلیل و بررسی کیفیت ضروری‌اند، درحالی‌که داده‌های پردازش‌شده برای تحلیل زیستی نهایی و تفسیر داده‌ها اهمیت دارند.

پ) پایگاه داده SRA (Sequence Read Archive) یکی از منابع اصلی ذخیره‌سازی داده‌های خام توالی‌یابی است که توسط NCBI نگهداری می‌شود. این پایگاه داده شامل توالی‌های خام از پروژه‌های RNA-Seq و سایر انواع توالی‌یابی است. برای دسترسی به داده‌های RNA-Seq در این پایگاه، می‌توان از شناسه‌های مربوط به پروژه‌ها یا نمونه‌ها استفاده کرد و داده‌ها را با ابزارهایی مانند SRA Toolkit دانلود کرد. این امکان برای محققان فراهم می‌شود تا داده‌های خام را جهت بازتحلیل یا استفاده در پروژه‌های جدید دریافت کنند.

سوالات عملی (۱۰۰ نمره)

۱. (۵۰ نمره) برای سوالات پاسخ به سوالات زیر به پایگاه داده GEO مراجعه فرمایید.

الف) داده‌ای به دلخواه انتخاب کنید و Accession Number آن را مشخص کنید (مثال: GSE4107). سپس با استفاده از GEO2R، آن را تحلیل کنید.

ب) مراحل جستجو و انتخاب داده را توضیح دهید.

ج) ژن‌های با تفاوت بیان معنی‌دار بین گروه کنترل و بیمار را شناسایی کنید.

د) نتایج را به‌صورت یک جدول شامل ژن‌های مهم و مقدار p-value ارائه دهید.

حل. ۱. جستجو و انتخاب داده در GEO
۲. در نوار جستجو، کلمه‌کلیدی مرتبط با تحقیق (مثلاً cancer microarray) را وارد کنید یا مستقیماً Accession Number داده دلخواه خود را وارد کنید. مثال: Accession Number = GSE۴۱۰۷ (داده‌های مربوط به سرطان روده).
۳. داده مورد نظر را انتخاب کنید و صفحه مربوط به آن را باز کنید. در این صفحه اطلاعاتی شامل نوع مطالعه، پلتفرم استفاده‌شده (مانند Affymetrix)، و لینک دانلود داده‌ها وجود دارد.
۴. استفاده از ابزار GEO۲R
۵. در صفحه داده، روی گزینه Analyze with GEO۲R کلیک کنید.
۶. گروه‌های نمونه‌ها (مانند Control و Case) را مشخص کنید:
۷. برای این کار از بخش Define groups استفاده کنید و نمونه‌ها را به دو گروه تقسیم کنید.
۸. تنظیمات پیش‌فرض آماری ابزار را بررسی کنید (برای شناسایی ژن‌های دیفرانسیلی بیان‌شده)
۹. اجرای تحلیل و استخراج ژن‌های دیفرانسیلی بیان‌شده (DEGs)
۱۰. پس از اجرای GEO۲R، جدولی تولید می‌شود که شامل اطلاعاتی مانند:

Gene Symbol: نام ژن
Log۲ Fold Change: میزان تغییر بیان ژن
p-value: سطح معنی‌داری آماری
۱۱. فیلترسازی کنید:
ژن‌هایی با |Log۲ Fold Change| > ۱ (افزایش یا کاهش بیان معنادار).
ژن‌هایی با p-value < ۰۵.۰ (سطح معنی‌داری).
۱۲. جدول خروجی را دانلود کنید و گزارش کنید.

۲. (۵۰ نمره) برای حل این سوال به نوتبوک تست‌های آماری در کوئرا مراجعه کنید.