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 | **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؟

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؟

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؟ بله این اتفاق اصلا شانسی نبوده 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:

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.

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

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؟

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؟

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

د) 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.

ه) 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.

و) 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.

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.

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