مقدمه‌ای بر بیوانفورماتیک **سری شمارهٔ ۲**

نیم‌سال دوم ۱۴۰۴ **موعد تحویل: ۳۰ آذر ۱۴۰۴**

طراحان: طه جلالی، رئوفه رضائی

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روزهای مشخص شده است.
- در طول ترم، برای هر تمرین می‌توانید تا ۵ روز تأخیر داشته باشید (به غیر از تمرین چهارم که تا ۳ روز تاخیر قابل قبول می‌باشد) و در مجموع ۱۲ روز تأخیر مجاز خواهید داشت. پس از گذشت ۵ روز از مهلت ارسال امکان ارسال پاسخ به اتمام میرسد و به ازای هر میزان از تأخیر غیر مجاز به صورت خطی از نمره آن تمرین کسر خواهد شد.
- در صورت وجود هرگونه ابهام میتوانید سوالات خود را از طریق صفحه آن تمرین در کوئرا مطرح کنید.
- لطفا پاسخ سوالات را در قالب یک فایل pdf به فرمت $STID\_HW۳.pdf$ در کوئرا آپلود کنید.

سؤال ۱.

غنی‌شدگی موتیف در نواحی پروموتر: آزمون دقیق جایگشت (Permutation Test)

در یک مطالعه‌ی RNA-seq (شرایط A در برابر کنترل)، شما ۳ ژن با بیان افزایشی (up-regulated) از میان ۶ ژن بیان‌شده شناسایی کرده‌اید. فرض شما این است که یک فاکتور رونویسی (TF-X) این پاسخ را از طریق موتیف شناخته‌شده‌ی خود، یعنی موتیف ۶ـ تایی `ACGTGC`، تنظیم می‌کند. توالی‌های ۳۰ جفت‌پایه‌ای از نواحی پروموتر برای این شش ژن در جدول زیر آمده است.

ژن	توالی ۳۰ جفت‌پایه‌ای پروموتر	بیان افزایشی؟
A	TTAAGGACGTGCTTCCGATGGAATTCGACA	خیر
B	TTGGAATGCCATTGTTGGAATCCATTTGGA	بله
C	GGTTCCGTAACGTGCGGATTAACCTGGAAT	بله
D	GCTTCTGGAATTTGCAATGGTTAACCAATT	خیر
E	CCATGGAAACGTGCTTAGGCTTAACGATGTT	بله
F	CGGTTTGAATGCCATTGTTGGAACCATTTG	خیر

آیا فراوانی وقوع `ACGTGC` در پروموترِ ژن‌های با بیان افزایشی بیش از حد انتظار است؟ فرض صفر و فرض جایگزین را دقیقا تعریف کرده و از یک آزمون دقیق جایگشت استفاده کنید و p-value یک‌طرفه را برای غنی‌شدگی محاسبه کنید. سطح معناداری را $\alpha = ۰/۰۵$ در نظر بگیرید.

سؤال ۲.

تحلیل بیان ژن و محاسبه‌یp-value تعدیل‌شده

در یک مطالعه‌ی تحلیل بیان ژن، شش ژن برای تفاوت بیان میان دو شرایط با استفاده از t-test دو نمونه‌ای بررسی شده‌اند. جدول زیر مقدار log2 Fold Change (تفاوت میانگین لگاریتمیِ بیان، حالت تیمارشده در برابر کنترل) و مقدار p-value خام هر آزمون را نشان می‌دهد.

ژن	log2FC	p-value خام
A	+1.2	0.004
B	+0.9	0.011
C	+0.7	0.021
D	+0.5	0.039
E	+0.3	0.081
F	-0.4	0.250

تعداد آزمون‌ها $m = ۶$ است. تمام محاسبات را به‌صورت دستی انجام دهید و مراحل میانی را نشان دهید.

۱. توضیح دهید که چرا نیاز به استفاده از p-value های تعدیل شده داریم و چرا مقادیر خام گمراه کننده هستند.

۲. با استفاده از روش Bonferroni مقادیر تعدیل شده را محاسبه کنید و مشخص کنید کدام ژن‌ها در سطح معناداری $\alpha = ۰/۰۵$ معنی‌دار هستند؟

۳. با استفاده از روش Benjamini-Hochberg مقادیر تعدیل شده را محاسبه کنید و مشخص کنید کدام ژن‌ها در سطح معناداری $\alpha = ۰/۰۵$ معنی‌دار هستند؟

۴. توضیح دهید چرا روش BH در این مثال تعداد ژن‌های بیشتری را معنی‌دار تشخیص می‌دهد نسبت به بونفرونی. همچنین توضیح دهید هر روش چه خطایی را کنترل میکند.

سؤال ۳.

فرض کنید k و ℓ اعداد صحیح مثبت ($k, \ell \geq ۱$) هستند. در هر یک از موارد زیر، رشته‌ی S با الگوی مشخصی داده شده است. با فرض اینکه یک کاراکتر پایان رشته \$ (که از نظر ترتیب الفبایی از تمام حروف دیگر کوچکتر است) به انتهای S اضافه می‌شود (یعنی ورودی الگوریتم رشته‌ی $S\$$ است)، تبدیل Burrows-Wheeler Transform یا همان رشته‌ی BWT خروجی را به صورت دقیق بر حسب k و ℓ محاسبه کنید.

الف. رشته‌ی S به صورت $(AT)^k A$ است (برای مثال اگر $k = ۲$ باشد، رشته برابر $ATATA$ خواهد بود).

ب. رشته‌ی S به صورت $T^k A^\ell$ است (برای مثال اگر $k = ۳$ و $\ell = ۴$ باشد، رشته برابر $TTTAAAA$ خواهد بود).

ج. رشته‌ی S به صورت $A^k T A^k$ است (برای مثال اگر $k = ۳$ باشد، رشته برابر $AAATAAA$ خواهد بود).

سؤال ۴.

در یک مطالعه‌ی Microarray، اثر یک داروی جدید بر روی ۱۰ ژن کاندیدا بررسی شده است. نرم‌افزار آماری، مقادیر P-value خام حاصل از آزمون t را برای این ۱۰ ژن گزارش کرده است. داده‌های خام (P-values) در جدول زیر آمده است:

۰/۰۶	۰/۰۰۹	۰/۰۳۵	۰/۶۰	۰/۰۰۳	۰/۱۵	۰/۰۱۴	۰/۸۲	۰/۰۰۰۲	۰/۰۴۵

با در نظر گرفتن سطح معنی‌داری کلی $\alpha = ۰/۰۵$، به سوالات پاسخ دهید:

الف. روش Bonferroni: با توجه به تعداد آزمون‌ها ($m = ۱۰$)، ابتدا مقدار آستانه جدید را محاسبه کنید. بر اساس این معیار، کدام‌یک از مقادیر فوق همچنان از نظر آماری معنی‌دار محسوب می‌شوند؟ (مقادیر آنها را ذکر کنید).

ب. روش FDR / Benjamini-Hochberg: فرمول محاسبه مقدار بحرانی را برای هر رتبه k بنویسید. جدول محاسباتی را تشکیل داده و بزرگترین رتبه k را که شرط الگوریتم در آن صدق می‌کند، بیابید. لیست نهایی ژن‌های پذیرفته شده در این روش کدامند؟

ج. به طور خاص، وضعیت ژن با مقدار P-value=0.014 را در دو روش بالا بررسی کنید. چرا سرنوشت این ژن در دو روش متفاوت است؟ (توضیح دهید که هر روش بر کنترل چه نوع خطایی تمرکز دارد).

موفق باشید.