



مقدمه‌ای بر بیوانفورماتیک

پاییز ۱۴۰۳

استاد: علی شریفی زارچی

مسئول تمرین: امیررضا آرانپور

مهلت ارسال نهایی: ۱۸ آبان

تمرین اول

مهلت ارسال بدون تاخیر: ۱۴ آبان

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روزهای مشخص شده 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 تأخیرهای مجاز، می‌توانید با تاخیری ساعتی ۱ درصد تمرین خود را ارسال کنید.
- حتماً تمرین‌ها را بر اساس موارد ذکر شده 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 `[STD_ID]_HW1.pdf` آپلود کنید.
- گردآوردندگان تمرین: امیررضا آرانپور، رها قره‌داغی، عرفان محمدی، زهرا رجالی

سوالات نظری (۱۰۰ نمره)

۱. (۱۵ نمره) فرض کنید ژنی که مسئول تولید پروتئینی خاص در سلول‌های عضلانی است، دچار جهش شده است. این جهش باعث شده است که عضلات توانایی بازسازی پس از آسیب را از دست بدهند.
 - الف) توضیح دهید که چگونه جهش در یک ژن می‌تواند باعث کاهش یا توقف بازسازی سلول‌های عضلانی شود.
 - ب) فرض کنید که این جهش ژنی در یک کروموزوم X قرار دارد و بیماری حاصل از این جهش به صورت وابسته به جنسیت منتقل می‌شود. اگر یک دختر و یک پسر این جهش را از یکی از والدینشان به ارث ببرند، توضیح دهید که هر کدام از آن‌ها در چه وضعیتی خواهند بود.
۲. (۱۵ نمره) در فرآیند تکثیر DNA، آنزیم‌های متعددی به طور هماهنگ کار می‌کنند تا تکثیر دقیق و کامل انجام شود. حال فرض کنید که جهشی در ژن کدکننده DNA Polymerase رخ داده که دقت این آنزیم را در جایگزینی نوکلئوتیدهای صحیح کاهش داده است. علاوه بر این، DNA Helicase نیز به دلیل جهشی دیگر قادر به باز کردن دو رشته DNA با سرعت معمول نیست.
 - الف) جهش در DNA Polymerase ممکن است منجر به افزایش تعداد خطاهای تکثیری شود. توضیح دهید که این خطاها چگونه می‌توانند بر توالی ژنی و در نهایت بر عملکرد پروتئین‌های تولیدی تأثیر بگذارند.
 - ب) اگر جهش در DNA Helicase باعث شود که این آنزیم نتواند به موقع دو رشته DNA را باز کند، توضیح دهید که چگونه این امر باعث ایجاد اختلال در تکثیر DNA و مشکلات ساختاری در رشته‌های جدید DNA می‌شود.
۳. (۱۰ نمره) یک زیست‌شناس گونه‌ای از باکتری‌ها را کشف کرده که در محیط‌های با دمای بسیار بالا مانند چشمه‌های آب‌گرم زندگی می‌کنند. این باکتری‌ها توانایی تولید آنزیم‌هایی دارند که در دماهای بالا نیز فعالیت خود را حفظ می‌کنند. او به دنبال شناسایی ژن‌های مسئول تولید این آنزیم‌های مقاوم به حرارت است و از روش RT-PCR برای بررسی بیان ژن‌ها استفاده می‌کند. به نظر شما چه چالش‌هایی ممکن است در فرآیند جداسازی پروتئین‌های این باکتری وجود داشته باشد و چگونه می‌توان از پایداری آنزیم‌ها در طی مراحل جداسازی و خالص‌سازی اطمینان حاصل کرد؟

۴. (۱۵ نمره) شما دو توالی DNA از دو گونه مختلف دارید.

x = AGCTGAC
y = AGCAGTC

الف) Hamming Distance را برای این دو توالی محاسبه کنید.
ب) Edit Distance را برای این دو توالی محاسبه کنید. توضیح دهید که چگونه این فاصله می‌تواند اطلاعات بیشتری نسبت به Hamming Distance درباره تغییرات ژنتیکی بین توالی‌ها ارائه دهد.

۵. (۱۵ نمره) در بیولوژی و توالی‌های ژنتیکی، فاصله‌ها (gaps) به بخش‌هایی از توالی اشاره دارند که در یک توالی وجود دارند اما در توالی دیگر حذف شده‌اند یا وجود ندارند. این فاصله‌ها به دلیل تغییرات طبیعی مانند حذف (deletion)، درج (insertion) یا جهش (mutation) در فرآیندهای تکاملی ایجاد می‌شوند. در هم‌ترازی توالی‌ها، دو نوع جریمه به نام‌های open gap و extension gap برای کنترل فاصله‌ها استفاده می‌شود. جریمه open gap بالاتر است زیرا ایجاد یک فاصله جدید تغییر چشمگیری در ساختار توالی به وجود می‌آورد و اغلب به تغییرات تکاملی مهم‌تری مانند حذف یا درج اشاره دارد. از سوی دیگر، extension gap جریمه کمتری دارد چون صرفاً به گسترش یک فاصله موجود می‌پردازد و تغییری جدید ایجاد نمی‌کند. در نتیجه، برای جلوگیری از ایجاد فاصله‌های زیاد و بی‌رویه، جریمه open gap بیشتر است تا هم‌ترازی دقیق‌تری بین توالی‌ها برقرار شود. با استفاده از سیستم امتیازدهی داده شده، دو توالی زیر را هم‌تراز کنید.

x = GCAC
y = GCC

سیستم امتیازدهی:

match: ۲

mismatch: -۲

gap open: -۵

gap extend: -۱

۶. (۳۰ نمره) **MUSCLE** معرفی شده در سال ۲۰۰۴ یک الگوریتم قدرتمند و مشهور برای هم‌ترازی چند توالی است که با هدف دستیابی به دقت بالا و سرعت بالا طراحی شده است. با توجه به مقاله آن به سوالات زیر پاسخ دهید.

الف) ایده اصلی MUSCLE چیست؟

ب) hits در این مقاله به چه معنا است؟

ج) مراحل اصلی MUSCLE چیست و هر مرحله چگونه به بهبود هم‌ترازی نهایی کمک می‌کند؟

د) روش k-mer در MUSCLE چگونه به محاسبه سریع شباهت جفتی توالی‌ها کمک می‌کند؟

ه) MUSCLE چگونه در مرحله سوم (refinement) (stage) هم‌ترازی را بهبود می‌بخشد؟ چرا این مرحله در دستیابی به دقت بیشتر اهمیت دارد؟

و) چه ویژگی‌هایی از MUSCLE باعث می‌شود که این الگوریتم برای مجموعه داده‌های بزرگ با تعداد زیادی از توالی‌ها مناسب باشد؟

۱. (۱۰۰ نمره) برای سوالات عملی به *quera* و سایت *rosalind* مراجعه کنید..

- (۱) (۶۰ نمره) پس از عضویت در کلاس تشکیل شده در *rosalind* به سوالات پاسخ دهید. برای این سوالات می‌توانید از هر زبان برنامه‌نویسی‌ای استفاده کنید.
- (۲) (۴۰ نمره) برای تمرین عملی دوم به *quera* مراجعه کنید. برای این تمرین می‌توانید از هر یک از زبان‌های *python*، *Java*، *C* و *C++* استفاده کنید.