دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

مقدمه‌ای بر بیوانفورماتیک

پاییز ۱۴۰۳

استاد: علی شریفی زارچی
مسئول تمرین: ساجده فدائی

مهلت ارسال بدون تاخیر: ۳ آذر | **تمرین دوم** | مهلت ارسال نهایی: ۷ آذر

- 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 $HW۲\_[STD\_ID].pdf$ آپلود کنید.
- گردآورندگان تمرین: ساجده فدائی، نیکان واسعی، محمد مولوی، امیرحسین علیشاهی

سوالات نظری (۱۰۰ نمره)

۱. (۱۵ نمره) در تکنولوژی‌های توالی‌یابی نسل جدید، (NGS) تعیین توالی بخش‌های تکراری ژنوم چالش‌های خاصی به همراه دارد. به سؤالات زیر پاسخ دهید:

الف) چرا توالی‌یابی بخش‌های تکراری ژنوم دشوار است و چگونه می‌توان این چالش‌ها را کاهش داد؟

ب) با توجه به تکنیک‌های موجود، چگونه از استراتژی‌های جفت‌ـخوانشی (paired-end) برای بهبود دقت توالی‌یابی در این بخش‌های تکراری استفاده می‌شود؟

۲. (۱۵ نمره) دانشمندان در حین تحقیقات خود در شناسایی ژن‌های عجیب دو گونه، توانستند readهای مربوط به آن‌ها را بدست آورند و در فایل‌های جداگانه ذخیره کنند. اما متاسفانه به دلیل اشتباه یکی از دانشمندان، این فایل‌ها با یک‌دیگر ادغام شدند و درون یک فایل قرار گرفتند و بازسازی ژن‌ها را سخت‌تر کردند. با توجه به readهای مخلوط شده، سعی کنید دو تا ژن اولیه را بازسازی کنید.

AAGT - TCTT - AGTA - CCAA - GTAG - ATCT - TAGG - ACCA - AGGA - GGAT - GATT - ATTT - ACTC - CTCG - TCGC - TTTG - GCAT - CATC - CTTA - CGCA - TTAC - TACC

۳. (۳۰ نمره) فرض کنید یک متد جدید برای alignment Multiple که قطعه کد آن در زیر معرفی شده است به شما داده شده، با این قطعه کد و جدول امتیازات زیر چهار رشته داده شده را با هم تراز کنید. (توجه داشته باشید که حتما جداول همترازی را به صورت کامل رسم کنید و صرفا نمایش خروجی هم‌تراز شده کامل نمی‌باشد و نمره‌ای به آن تعلق نمیگیرد)

	-	A	T	C	G
-	۰	-۱	-۱	-۱	-۱
A	-۱	۲	۱	۰	۰
T	-۱	۱	۲	-۱	۱
C	-۱	۰	-۱	۳	۲
G	-۱	۰	۱	۲	۳

Multiple Sequence Alignment Algorithm

Require: A set S of sequences
Ensure: A multiple alignment of M with sum of pair distances at most twice that of the optimal alignment of S

۱ Find $D(S_i, S_j)$ for all i, j.
۲ Find the center sequence S_c which minimizes $\sum_{i=1}^{k} D(S_c, S_i)$.
۳ **for** each $S_i \in S - \{S_c\}$
۴ Choose an optimal alignment between S_c and S_i.
۵ Introduce spaces into S_c so that the multiple alignment $\mathcal{M}$ satisfies the alignments found in Step ۳.

maximize similarity=minimize distance

رشته‌های داده شده به صورت زیر هستند:

S
ACGCTGAACG
ACTCGGAGC
CTGGAATCT
GCTAGGACC

۴. (۱۰ نمره) امروزه روش هم‌ترازی ساختاری (structural alignment) برای هم‌ترازی ماکرومولکول‌هایی مانند پروتئین‌ها به کار می‌روند.

الف) در مورد این روش تحقیق کنید و مراحل اصلی و کلیدی را به ترتیب و بطور کامل توضیح دهید.

ب) دو مورد از الگوریتم‌های کاربردی در این روش را نام برده و در خصوص آن‌ها توضیح دهید.

ج) در خصوص تفاوت دو روش هم‌ترازی مبتنی بر توالی و هم‌ترازی ساختاری توضیح دهید.

د) چگونه روش‌های تراز ساختاری می‌توانند هم‌ترازی توالی‌های چندگانه مبتنی بر توالی سنتی را برای استنباط روابط تکاملی و عملکردی در پروتئین‌ها تکمیل کنند، در مورد چالش‌ها و محدودیت‌های ادغام داده‌های ساختاری در MSA توضیح دهید.

۵. (۱۵ نمره) علاوه بر الگوریتمهای سنتی ساخت درخت فیلوژنی، از روشهای دیگری هم مانند Maximum Likelihood و یا استنتاج بیزی (Bayesian Inference) در ساخت این درختها استفاده میشوند. به پرسشهای زیر در رابطه با هر یک از این الگوریتمها پاسخ دهید:

الف) نحوه کار این الگوریتم چگونه است؟

ب) آن را با الگوریتم Neighbor-Joining مقایسه کنید و بررسی کنید که در چه مواقعی بهتر است از این الگوریتم استفاده کنیم.

ج) دو الگوریتم استنتاج بیزی و Maximum-Likelihood را با هم مقایسه کنید و برتری هر یک را نسبت به دیگری بیان کنید.

۶. (۱۵ نمره) الگوریتم UPGMA را برای ماتریس فاصله زیر پیاده سازی کنید.

	A	B	C	D	E	F	G
A							
B	5						
C	9	10					
D	9	10	8				
E	8	7	7	3			
F	7	4	10	9	6		
G	12	11	6	5	4	9	

شکل ۱: ماتریس فاصله

سوالات عملی (۱۰۰ نمره)

۱. (۱۰۰ نمره) برای سوالات عملی به $quera$ مراجعه کنید..

(۱) (۷۰ نمره) برای تمرین عملی اول و دوم به $quera$ مراجعه کنید. برای این تمرین می‌توانید از هر یک از زبان‌های $python$، $Java$، C و $C++$ استفاده کنید.

(۲) (۳۰ نمره) برای تمرین عملی دوم یک فایل در قالب $jupyter\ notebook$ در اختیار شما قرار گرفته است که بایستی آن را دانلود و تمامی بخش‌های خواسته شده را به صورت کامل و بدون خطا اجرا نموده و آن را در محل تعیین شده آپلود نمایید.