به نام خدا

- **گزارش کاملی از راه حل خود بنویسید. این گزارش باید شامل توضیحات مربوط به تمام بخش‌های کدها، کتابخانه ها و اسکریپت های استفاده شده در هر مرحله و نتیجه گیری نهایی باشد. دقت کنید معیار ارزیابی شما فقط گزارش شما است اما علاوه بر گزارش اسکریپت‌ها و کدها را به صورت ضمیمه در کنار گزارش قرار دهید تا در صورت نیاز و یا بروز مشکلی بررسی شود.**
- **فایل ضمیمه به صورت یک فایل فشرده و فرمت نام‌گذاری آن به صورت CG_PR_student name_student id باشد.**
- **برای آشنایی بیشتر و بالا رفتن اطلاعات تجربی و عملی خود شما باید از پکیج یا بسته نرم افزاری متفاوت برای پاسخ دهی به سوالات دو و سه استفاده کنید. در انتخاب نرم افزار و پکیج کاملا آزاد هستید (پکیج‌های موجود در R یا پایتون یا نرم افزاری مشابه mega و یا هر منبع دیگر...) اما حتما اطلاعات مربوط به بسته استفاده شده و مراحل انجام گام به گام کار را به طور کامل در گزارش خود بیاورید**

سوال یک: در این تمرین قصد داریم با استفاده از نرم افزار mega اقدام به رسم درخت فیلوژنی و تحلیل آن کنیم:

https://www.megasoftware.net/

الف) داده: از فایل ضمیمه data.fasta که شامل رشته ریبوزومی از ۱۰ گونه یا سویه مختلف باکتریایی می‌باشد استفاده کنید.

ب) alignment : برای رسم درخت فیلوژنی ابتدا نیاز است نمونه‌های خود را align کنیم. از قسمت align استفاده کرده و با الگوریتم muscle داده ها را align کنید. (اینکار را با استفاده از گزینه without codon انجام دهید و توضیح دهید که استفاده از codon alignment چه مزایا و معایبی در رسم درخت تکاملی می‌تواند داشته باشد)

ج) شرط استفاده از الگوریتم neighbor-joining داشتن میانگین فاصله jc(jukes-cantor) کمتر از ۱ برای تمام جفت فاصله ها است.علت این موضوع را بنویسید. همچنین این مقدار را به کمک قسمت distance محاسبه کرده و از مناسب بودن شرایط داده خود مطمئن شوید و نتیجه را در گزارش بیاورید.

د) درخت فیلوژنی را با الگوریتم‌های Maximum likelihood,neighbor-joining,UPGMA,maximum parsimony رسم کنید و در گزارش خود بیاورید.

- به صورت خلاصه مراحل الگوریتم ها را شرح داده و سرعت محاسباتی این الگوریتم ها را مقایسه کنید.
- همچنین در صورت وجود تفاوت در درخت خروجی الگوریتم ها علت بروز تفاوت را بررسی و بیان کنید.
- با توجه به اینکه داده‌ها لیبل خورده هستند و برای هر داده گونه و سویه آن روی درخت مشخص است. تحلیل کنید آیا درخت درست تخمین زده شده است و فواصل روی درخت با گونه و سویه‌های باکتری‌ها همخوانی دارند یا خیر.

سوال دو : پروژه هزار ژنوم گردآورنده بزرگترین مجموعه دادگان ژنوم انسان با دسترسی عمومی است. فاز ۳ این پروژه شامل ۲۵۰۴نمونه از ۲۶جمعیت مختلف است.

برای اطلاعات بیشتر در مورد این پروژه به سایت هزارژنوم مراجعه کنید:

https://www.internationalgenome.org/data-portal/data-collection/phase-3

برای این سوال از داده میتوکندری هر یک از ۲۶ جمعیت یک نمونه تصادفی انتخاب کنید و پاسخ سوال ها را بیابید.

فایل میتوکندری همه نمونه ها:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chrMT.phase3_callmom-v0_4.20130502.genotypes.vcf.gz

۱.مطلوبست رسم درخت اجدادی برای رشته ها

۲.تخمین زمان MRCA

سوال سه: به سایت ncbi بروید و ۲۰ نمونه از رشته های covid19 که در بازه های ۳ ماهه استخراج شده باشند را دانلود کنید

۱.مطلوبست درخت اجدادی برای رشته ها

۲.تخمین زمان MR