موضوع کلی این پروژه، مقایسه ژنوم‌هاست. شما می‌توانید یکی از دو پروژه الف و یا ب را به عنوان پروژه پایانی خود انتخاب کنید. هر یک از این موضوعات ظرفیت مشخصی دارند و تنها تعداد محدودی از گروه‌ها می‌توانند هر یک از این دو را انتخاب کنند. شما فرصت دارید تا تاریخ ۱۸ اردیبهشت موضوع انتخابی خود و اسامی اعضای گروه را در این لینک مشخص کنید. موضوع به هیچ وجه قابل تغییر نیست.

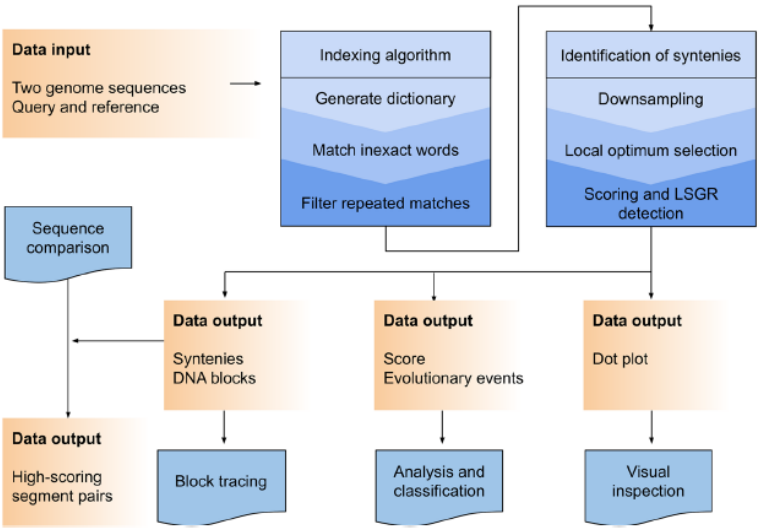
نکته ۱: شما می‌توانید به صورت گروهی برای ارائه پروژه اقدام کنید. گروه‌های شما می‌تواند حداکثر دو نفره باشد.

نکته ۲: هر یک از پروژه‌ها شامل فازهای مختلفی است که مهلت ارسال مخصوص به خود را دارند. جدول زمانبندی مهلت ارسال فازهای مختلف پروژه در جدول زیر مشخص شده است. به ازای هر روز دیرکرد ارسال هر بخش، ۲۰٪ از نمره‌ی آن مرحله کسر می‌گردد.

فاز پروژه	مهلت ارسال	درصد نمره	توضیحات
فاز اول	پایان روز ۱۰ خرداد	٪۳۰	زمان‌بندی دقیق ارائه‌ی هر یک از گروه‌ها اطلاع‌رسانی خواهد شد
فاز دوم	پایان روز ۲۰ تیر	٪۳۵	
فاز سوم	پایان روز ۱۰ مرداد	٪۳۵	

الف. مقایسه بسیار سریع ژنوم‌های بزرگ مقیاس

این پروژه بر اساس مقاله Ultra-fast genome comparison for large-scale genomic experiments طراحی شده است. روش CHROMEISTER که توسط این مقاله ارائه شده است، یک روش ترکیبی است که آن را قادر به کاهش فضای جستجو برای مقایسه‌ی ژنوم‌های بزرگ می‌کند و به این ترتیب در زمان اجرا و حافظه صرفه‌جویی می‌کند. هدف این روش مقایسه ژنوم‌ها و گزارش امتیاز برای این مقایسه است. اما خروجی‌های میانی ارزشمندی چون نمودار شبه dot plot نیز گزارش می‌کند. مراحل کلی این روش و خروجی‌های آن در شکل ۱ قابل مشاهده است.



شکل ۱ مراحل اجرا و خروجی‌های روش CHROMEISTER

فاز اول:

در فاز اول لازم است این مقاله را خوانده و ارائه دهید. ارائه شما باید شامل گزارش method این مقاله و نحوه‌ی ارزیابی آن باشد.

نکته ۱: جدول زمانبندی ارائه‌ها پس از مشخص شدن گروه‌بندی‌ها و تخصیص موضوعات پروژه، اعلام می‌شود.

فاز دوم:

در این فاز، شما باید ابزاری مشابه ابزار CHROMEISTER را ارائه دهید.

ابزار توسعه یافته توسط شما باید تمامی متریک‌ها و نمودارهایی که روش CHROMEISTER گزارش می‌دهد را شامل شود. همچنین به انتخاب خود یکی از data setهای مقاله مذکور را انتخاب کرده و **ابتدا آن را از طریق ایمیل saeedeh.akbarira@gmail.com اطلاع داده و تأییدیه استفاده از آن را دریافت کنید** و سپس آزمایشی مشابه با مقاله با ابزار خود بر روی آن‌ها انجام داده و صحت عملکرد ابزار خود را نشان دهید. البته بعد از انتخاب data set خود، ابتدا ایمیلی به زده و تأییدیه استفاده از آن را دریافت کنید.

فاز سوم:

مجموع دو رقم کم ارزش شماره دانشجویی خود و همگروهیتان را در نظر بگیرید (در صورت انجام پروژه به صورت انفرادی، دو رقم کم ارزش شماره دانشجویی خود را در نظر بگیرید). گزینه برابر با باقیمانده این عدد بر ۳، مرحله بعدی کار شماست. با انجام مورد مربوط به خود، دقت، زمان و حافظه‌ی مورد نیاز برای این حالت را با فاز دوم مقایسه کنید.

۰. به جای k-mer به طول ۳۲ که توسط مقاله انتخاب شده است، مقادیر ۸، ۱۰ و ۱۲ را به عنوان طول k-mer انتخاب کنید و برای فیلترکردن ابتدا کمینه تعداد تکرار را از تمام مقادیر کم کرده و سپس کمینه مقادیر جدید را به جای یک بار تکرار انتخاب کنید (عدد تکرار را گزارش کنید). سپس آزمون فاز دوم را برروی این حالات بررسی کنید.

۱. به جای توابع F مقاله به عنوان تابع subsampling، ۱۱۱۱**********۱۱۱۱**********۱۱۱۱********** و ۱۱**********۱۱**********۱۱**********۱۱ و ۱۱***۱***۱۱***۱***۱۱***۱***۱۱***۱***۱۱***۱***۱۱***۱** را به عنوان توابع subsampling انتخاب کنید و برای فیلترکردن ابتدا کمینه تعداد تکرار را از تمام مقادیر کم کرده و سپس کمینه مقادیر جدید را به جای یک بار تکرار انتخاب کنید (عدد تکرار را گزارش کنید). سپس آزمون فاز دوم را برروی این حالات بررسی کنید. (۱ها نشان‌دهنده‌ی نوکلئوتیدهای match و * برای حالت mismatch هستند).

۲. به جای توابع F مقاله به عنوان تابع subsampling، ۱۱*۱***۱*******۱***۱*۱۱۱۱*******۱***۱*۱۱ و ۱*۱*۱***۱*۱*۱***۱*۱*۱***۱*۱*۱*** و *۱۱**۱***۱۱**۱***۱۱**۱***۱۱**۱***۱۱**۱** را به عنوان توابع subsampling انتخاب کنید و برای فیلترکردن ابتدا کمینه تعداد تکرار را از تمام مقادیر کم کرده و سپس کمینه مقادیر جدید را به جای یک بار تکرار انتخاب کنید (عدد تکرار را گزارش کنید). سپس آزمون فاز دوم را برروی این حالات بررسی کنید. (۱ها نشان‌دهنده‌ی نوکلئوتیدهای match و * برای حالت mismatch هستند).

ب. مقایسه مستقل از هم‌ترازی

این پروژه بر اساس مقاله CAFE accelerated Alignment-Free sequence analysis طراحی شده است. این ابزار ۲۸ روش مقایسه مستقل از هم‌ترازی را گردآوری کرده است.

فاز اول:

در فاز اول لازم است این مقاله را خوانده و ارائه دهید. ارائه شما باید شامل گزارشی از سه دسته روش مقایسه مستقل از هم‌ترازی ذکر شده در این مقاله باشد.

نکته ۱: جدول زمانبندی ارائه‌ها پس از مشخص شدن گروه‌بندی‌ها و تخصیص موضوعات پروژه، اعلام می‌شود.

فاز دوم:

در این فاز، شما باید یک معیار میزان شباهت دو ژنوم را پیاده‌سازی و ارزیابی کنید.

۱. تابعی طراحی کنید که قابلیت تقسیم یک ژنوم به بلوک‌هایی با شرایط زیر را داشته باشد:
   ۱. تعداد نوکلئوتید داخل هر بلوک به عنوان ورودی تابع قابل تغییر باشد.
   ۲. همپوشانی بلوک‌ها به عنوان ورودی تابع قابل تغییر باشد (به این معنی که دادن عدد صفر یعنی دو بلوک با هم هیچ همپوشانی نداشته باشند و هر عدد صحیح دیگر n یعنی بلوک iام و i+1ام، n نوکلئوتید مشترک داشته باشند.
   ۳. خروجی تابع باید یک فایل با فرمت fasta باشد که هر بلوک با یک کامنت از نوع این فایل جدا شده باشد. همچنین تعداد بلوک باید در کامنت اول این فایل نوشته شود.
۲. دو مجموعه ژنوم در اختیار شما قرار گرفته است (آنفولانزا و ژنوم میتوکندری پستانداران، دوستانی که مجموع دو رقم کم ارزش شماره دانشجویی خود و همگروهیشان فرد است آنفولانزا و اگر زوج است میتوکندری را انتخاب کنند). پس از اعمال تابع مرحله ۱ خود بر روی داده انتخابی خود، طبق جدول ۱، تعداد k-merها به طول ۳، ۵، ۷، ۹، ۱۱، ۲۴ و ۳۲ را برای این مجموعه‌ها به‌دست آورید. صحت عملکرد ابزار خود را تا به اینجای کار نشان دهید.

حالت	تعداد نوکلئوتید داخل بلوک	میزان همپوشانی	مجموعه داده
۱	۲۰۰	۰	آنفولانزا
۲	۲۰۰	طول k-mer	آنفولانزا
۳	۱۰۰۰	۰	میتوکندری
۴	۱۰۰۰	طول k-mer	میتوکندری

۳. دو رقم کم ارزش شماره دانشجویی خود را در نظر بگیرید، در صورت انجام کار به صورت گروهی مجموع دو رقم کم ارزش شماره دانشجویی خود و همگروهیتان را در نظر بگیرید. گزینه برابر با باقیمانده این عدد بر ۴، مرحله بعدی کار شماست. خروجی فیلتر شده‌ای طبق مرحله‌ی خود ارائه دهید.

۰. فیلتر کردن شمارش k-merها بر اساس بودن و نبودن (هر k-merای که یک یا بیشتر تکرار دارد مقدار ۱ و در غیر این صورت ۰ می‌گیرد)

۱. نگاشتن شمارش k-merها به بازه ۰ الی ۲۵۵ (بیشینه تکرار k-mer در کل ژنوم به عنوان ۲۵۵ و ۰ به عنوان کمینه مقدار در نظر گرفته می‌شود و شمارش k-merها به این بازه تبدیل می‌شوند. باید اعداد اعشاری به سمت کمتر گرد کنید).

۲. نگاشتن شمارش k-merهای بیشتر از ۲۰۰ به بازه ۲۰۰ الی ۲۵۵ (اعداد ۰ الی ۲۰۰ در تعداد k-merها را خودشان حفظ کرده و بیشینه تکرار k-mer در کل ژنوم به عنوان ۲۵۵ و ۲۰۰ به عنوان ۲۰۰ در نظر گرفته می‌شود و شمارش k-merها به این بازه تبدیل می‌شوند. باید اعداد اعشاری به سمت کمتر گرد کنید).

۳. ادغام کردن kmerهایی که با شروع از ابتدای آن‌ها، یکی درمیان نوکلئوتیدهای مشابه دارند (به این معنی که از الگوی ۱*۱*۱... پیروی می کنند. ۱ها نشان‌دهنده‌ی نوکلئوتیدهای match و * برای حالت mismatch هستند).

فاز سوم:

۱. دو رقم کم ارزش شماره دانشجویی خود را در نظر بگیرید، در صورت انجام کار به صورت گروهی مجموع دو رقم کم ارزش شماره دانشجویی خود و همگروهیتان را در نظر بگیرید. گزینه برابر با باقیمانده این عدد بر ۴، متریک مقایسه‌ی شما خواهد بود:

۰. $\sum_{i=1}^{4^k} kmer_{1,i} \times kmer_{2,i}$ دو گونه

۱. $\frac{1}{4^k}\left(\sum_{i=1}^{4^k}\frac{\left(kmer_{1,i}-mean(kmer_1)\right).\left(kmer_{2,i}-mean(kmer_2)\right)}{\sqrt{var(kmer_1)var(kmer_2)}}\right)$ دو گونه

۲. $\sum_{i=1}^{4^k}|kmer_{1,i}-kmer_{2,i}|$ دو گونه

۳. $\sum_{i=1}^{4^k}\frac{|kmer_{1,i}-kmer_{2,i}|}{kmer_{1,i}+kmer_{2,i}}$ دو گونه

متریک خود را برای مقایسه بلوک به بلوک هر دو گونه از مجموعه داده خود محاسبه کنید و در ماتریس مشابه dotplot به عنوان خروجی گزارش دهید.

۲. مجموع تمام مقادیر ماتریس مشابه dot plot بخش قبل را به عنوان معیار شباهت (برای متریک ۰ و ۱) و به عنوان معیار تفاوت (برای متریک ۲ و ۳) دو گونه در نظر گرفته و ماتریس شباهت (یا تفاوت) کل data set خود را گزارش دهید. همچنین به کمک ابزار MEGA درخت فیلوژنی به روش UPGMA برای آن رسم کرده و گزارش دهید (برای متریک ۰ و ۱ کل ماتریس را از مقدار بیشینه ماتریس‌ها کم کنید تا ماتریس تفاوت شود).

توضیحات

۱. برای فاز اول شما باید گزارشی از ارائه خود را با نام PRJ_Phase1_[StudentID1]_[StudentID2] ارسال کنید.
۲. برای فاز دوم و سوم متناسب با موارد خواسته شده در هر مرحله، شما باید کد ابزار پیاده‌سازی شده، نتایج، گزارشی شامل نحوه‌ی پیاده‌سازی و نتایج و همچنین دستورالعملی از نحوه‌ی استفاده از کد (در قالب یک فایل به اسم README به زبان انگلیسی یا فارسی) را به صورت zip شده و با اسم PRJ_Phase[2 or 3]_[StudentID1]_[StudentID2] ارسال کنید.
۳. به تمیز بودن کد و داشتن commentهایی که موجب خوانایی کد شود نمره اضافه تعلق خواهد گرفت.
۴. اجرای حالاتی جز حالات در نظر گرفته شده در صورت پروژه برای هر کدام از آن‌ها که سبب بهبود دقت و یا سرعت و یا هردو آن‌ها گردد، نمره اضافه تا ۴۰٪ نمره پروژه تعلق می‌گیرد.
۵. برای توسعه ابزار خود می‌توانید از یکی از زبان‌های C، Java، Python، MATLAB و یا R استفاده کنید.
۶. شما در صورت نیاز، تنها مجاز به استفاده از ابزار آماده‌ی Jellyfish برای شمارش k-merها هستید که البته خروجی و روش استفاده از آن نیز باید در گزارش شما ذکر شود. استفاده از دیگر ابزارهای آماده جایز نیست.
۷. پروژه‌های خود را به ایمیل (saeedeh.akbarira@gmail.com) ارسال کنید. همچنین عنوان ایمیل را BIOALG_PRJ_Phase[1 or 2 or 3] قرار دهید.