به نام خدا

فاز ۲ پروژه‌ی الگوریتم‌های بیوانفورماتیک

نحوه‌ی پیاده سازی:

پیاده سازی در زبان پایتون انجام شده. نحوه‌ی اجرای الگوریتم از مقاله و همچنین کدهای زیر که متعلق به نویسنده‌ی مقاله است، بدست آمده:

- CHROMEISTER.c
- commonFunctions.c
- compute_score.R

فایل CHROMEISTER.py شامل کلاس CHROMEISTER است که پیاده سازی الگوریتم داخل آن انجام شده است.

فایل main.py برای اجرای برنامه از طریق console و پردازش آرگومان‌ها است که در ادامه توضیح داده خواهد شد.

فایل test.ipynb اجرای مرحله به مرحله و نتایج آن را نشان میدهد.

پکیج‌های مورد نیاز برای اجرای کد، در فایل requirements.txt آمده است.

نحوه‌ی اجرای کد:

برای اجرای الگوریتم از طریق console، باید از فایل main.py استفاده کنید. این فایل ابتدا آرگومان‌ها را پردازش کرده و سپس الگوریتم را با پارامترهای داده شده، اجرا میکند. آرگومان‌های ورودی به شرح زیر می‌باشند (وارد کردن آرگومان‌های قرمز اجباری می‌باشد. همچنین آرگومان‌های سبز با توجه به ایده‌های خودمان اضافه شده است):

- h-: نمایش پیغام راهنمایی در مورد آرگومان‌های ورودی
- --db: آدرس فایل fasta ژنوم Database. k-merهای یکتا و بدون همپوشانی از این ژنوم استخراج می‌شود. در نمودار dot-plot این ژنوم بر روی محور X قرار می‌گیرد.
- --query: آدرس فایل fasta ژنوم Query. k-merها و مکمل معکوس (Reverse Complement) آن‌ها از این ژنوم استخراج شده و در k-merهای دیتابیس به صورت غیر دقیق به دنبال آن‌ها گشته می‌شود. در نمودار dot-plot این ژنوم بر روی محور Y قرار می‌گیرد.
- --kmer-len: طول k-merها.
- --kmer-key-len: طول کلید k-merها که به صورت دقیق مقایسه می‌شود.
- --z: مقدار z در محاسبه‌ی hash.
- --dimension: اندازه‌ی طول و عرض ماتریس برخورد و dot-plot.
- --out-dir: آدرس فولدر جهت ذخیره‌ی خروجی‌ها.
- --diag-len: میزان گسترش نقاط در قطرها.
- --neighbour-dist: فاصله‌ی نقاط تا نقطه‌ی ماکسیمم در سطرها و ستون‌ها که در صورت داشتن مقدار، حذف نمی‌شوند. (در قسمت ایده‌ها توضیح داده خواهد شد)
- --kernel-width: طول و عرض کرنلی که در آن بررسی می‌شود یک نقطه باید به عنوان نقطه‌ی پرت، حذف شود یا بماند برابر دو برابر این مقدار به اضافه‌ی یک می‌باشد. (در قسمت ایده‌ها توضیح داده خواهد شد)
- --dist-th: آستانه فاصله‌ی استفاده شده در محاسبه امتیاز.
- --sampling-value: ضریب کوچک کردن dot-plot برای اینکه HSPها بهتر پیدا شوند.

- diag-separation--: آستانه‌ی فاصله از قطر اصلی برای تشخیص نوع رویدادها (Type of Events).
- hsp-th--: حداقل اندازه‌ی HSP
- verbose--: در صورتی که ۱ باشد، اطلاعاتی در هنگام اجرا نیز نمایش داده می‌شود.

ایده‌ها:

ایده‌ها در فایل CHROMEISTER.py و با کلمه‌ی Start New Idea مشخص شده‌اند. در زیر در مورد آن‌ها توضیحاتی میدهیم:

۱. خط ۲۵۶: در هنگام تمیز کردن Hit Matrix و بدست آوردن Dot-Plot، در مقاله تنها نقاط ماکسیمم در هر سطر و ستون نگه داشته می‌شود که باعث نازک شدن خطوط می‌شود. **ایده‌ی ما** این بود که تا یک فاصله‌ای از نقاط ماکسیمم در سطر و ستون‌ها (که با neighbour-dist-- مشخص می‌شود) نیز در صورتی که در Hit Matrix مقدار بزرگتر از صفر داشته باشند، در Dot-Plot نیز بیایند.

۲. خط ۳۱۵: در هنگام تمیز کردن Hit Matrix و بدست آوردن Dot-Plot، در مقاله تنها نقاطی که هیچ همسایه‌ای ندارند حذف می‌شوند. **ایده‌ی ما** این بود که کرنل تشخیص نقطه‌ی تنها را بزرگتر کنیم، همچنین مقدار آستانه نیز با توجه به اندازه‌ی کرنل مشخص می‌شود. (اندازه‌ی کرنل با kernel-width-- مشخص می‌شود)

۳. خط ۳۶۷: در هنگام پیدا کردن HSPها، در کد مقاله به صورت همزمان قطرهای اصلی و فرعی جستجو میشد که چون الگوریتم نقاطی را که مشاهده می‌کرد، پاک میکرد تا در مراحل بعدی دوباره به آن‌ها برنخورد، اگرHSPها از میزانی به هم نزدیکتر بودند، تنها یکی از آن‌ها پیدا میشد. **ایده‌ی ما** این بود که HSPهای قطر اصلی و فرعی را به صورت جدا، بررسی کنیم.

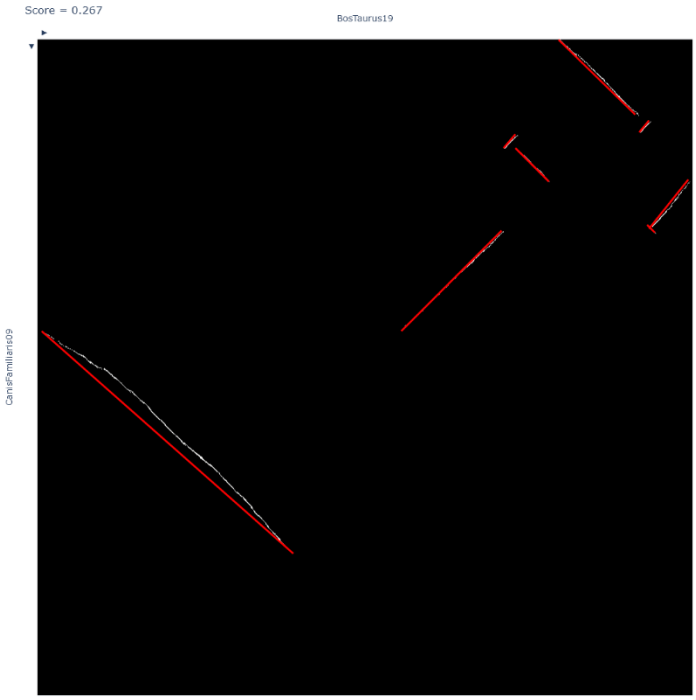
۴. خط ۵۷۱: در هنگام پیدا کردن HSPها، در کد مقاله نقطه‌ی انتهایی که برای سطر انتخاب میشد، در اکثر موارد به هیچ نقطه‌ی مقدار داری در dot-plot اشاره نمیکرد.

ایده‌ی ما این بود که آخرین سطری که در آن مقدار وجود دارد را به عنوان نقطه‌ی انتهایی HSP اعلام کنیم.

۵. خط ۶۱۴: در هنگام پیدا کردن نوع HSPها، در کد مقاله بررسی وجود HSP روی قطر اصلی به نحوی بود که inversion اگر مقداری طولانی بود به صورت inverted transposition تشخیص داده میشد. **ایده‌ی ما** این بود که نقطه‌ی وسط HSP را بررسی کنیم.

نتایج:

نتایج به صورت کامل برای z=2 و z=4 همراه کد فرستاده شده است. در زیر تنها به dot-plot آن‌ها نگاهی می‌اندازیم (کروموزوم ۱۹ BosTaurus و کروموزوم ۹ CanisFamiliaris):



نمودار dot-plot و LSGRها برای z=4

نمودار dot-plot و LSGRها برای z=2