به نام خدا

پروژه اول: داده ی متنی

پیدا کردن داده ساختار مدنظر

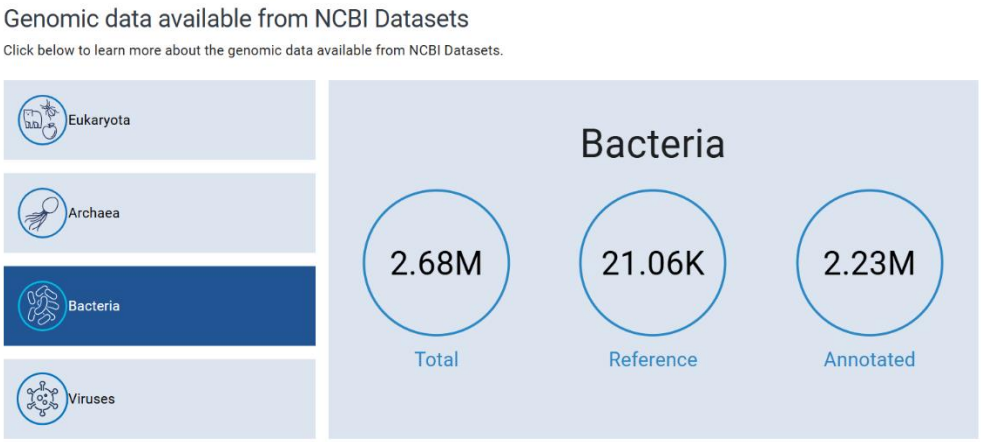
در ابتدا نظرم این بود که دیتاستی پیدا کنم که motifهای حفاظت شده را در پروتیین های housekeeping میان گونه ای شناسایی کند. برای مثال پروتیین actin.

پروتئین housekeeping یعنی پروتئینی که دائما بیان میشود(ژنش روشن است).

برای اینکار حدود 11 ساعت در سایت های زیر مشغول سرچ و دانلود دیتاست های با حجم بالا شدم:

ncbi.nlm.nih.gov	blast.ncbi.nlm.nih.gov	ukbiobank.ac.uk	uniprot.org
ebi.ac.uk	kaggle.com	(سایتهای دیگر-بیش از 8 سایت- زیرمجموعه NCBI هستند.)	

در دیتابیس های مربوط به زیست شناسی، داده ای که میخواستم (توالی نوکلئوتیدی یا پروتئینی، یا ریکورد بیماری یا mutationدر توالی ژنتیکی) به تعداد 1M پیدا نشد. پس تصمیم گرفتم به سراغ باکتری ها بروم.
چون گونه های کشف شده باکتریایی2.68M هستند:



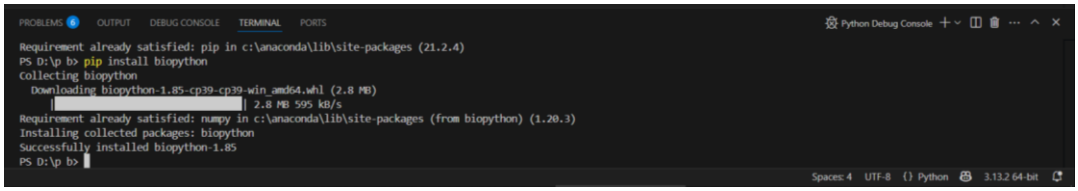
موضوعی که انتخاب کردم: مقایسه توالی ATP Synthase در 1 میلیون گونه باکتریایی.
پروتئین ATP Synthase یک پروتئین universal است. یعنی تقریبا در تمام موجودات زنده از ابتدای ایجاد حیات، حفاظت شده و وجود دارد. پس گزینه خوبی برای پروژه من است.
دوباره با یک چالش روبه رو شدم:

در NCBI که تقریبا تمام Dataset ها را دربرمیگیرد، توالی 1M گونه باکتریایی قابل دانلود درقالب یک فایل نبود. وارد کردن تک تک آنها بصورت دستی هم امکان پذیر نبود. پس تصمیم گرفتم بدنبال ابزاری باشم که بتواند تک تک توالی ها را استخراج کرده و در یک فایل ذخیره کند.

از copilot خواستم چنین ابزاری را معرفی کند.

جوابی که داد یک کد آماده بود که باید وارد biopython میکردم.

تفاوت python و biopython را پرسیدم و خواستم به من یاد بدهد چطور در VS Code بایوپایتون را نصب کنم:



بعد از حدود 2 ساعت ادیت و اصلاح کد مربوط به دریافت 1 میلیون توالی، NCBI من را بلاک کرد!. خروجی تا قبل از بلاک شدن ایمیلم و ناقص ماندن اجرای کد:

✅ دریافت 10000 شناسه تا این لحظه.
✅ دریافت 20000 شناسه تا این لحظه.
✅ دریافت 30000 شناسه تا این لحظه.
✅ دریافت 40000 شناسه تا این لحظه.
✅ دریافت 50000 شناسه تا این لحظه.
✅ دریافت 60000 شناسه تا این لحظه.
✅ دریافت 70000 شناسه تا این لحظه.
✅ دریافت 80000 شناسه تا این لحظه.
✅ دریافت 90000 شناسه تا این لحظه.
✅ دریافت 100000 شناسه تا این لحظه.
✅ دریافت 110000 شناسه تا این لحظه.
✅ دریافت 120000 شناسه تا این لحظه.
✅ دریافت 130000 شناسه تا این لحظه.
✅ دریافت 140000 شناسه تا این لحظه.
✅ دریافت 150000 شناسه تا این لحظه.
✅ دریافت 160000 شناسه تا این لحظه.

✅ دریافت 490000 شناسه تا این لحظه.
✅ دریافت 500000 شناسه تا این لحظه.
✅ دریافت 510000 شناسه تا این لحظه.
✅ دریافت 520000 شناسه تا این لحظه.
✅ دریافت 530000 شناسه تا این لحظه.
✅ دریافت 540000 شناسه تا این لحظه.
✅ دریافت 550000 شناسه تا این لحظه.
✅ دریافت 560000 شناسه تا این لحظه.
✅ دریافت 563371 شناسه تا این لحظه.
❌ هیچ نتیجه‌ای یافت نشد! بررسی کن که محدودیت‌های NCBI رعایت شده باشن.
شروع دریافت توالی‌های کامل...

✅ دریافت 170000 شناسه تا این لحظه.
✅ دریافت 180000 شناسه تا این لحظه.
✅ دریافت 190000 شناسه تا این لحظه.
✅ دریافت 200000 شناسه تا این لحظه.
✅ دریافت 210000 شناسه تا این لحظه.
✅ دریافت 220000 شناسه تا این لحظه.
✅ دریافت 230000 شناسه تا این لحظه.
✅ دریافت 240000 شناسه تا این لحظه.
✅ دریافت 250000 شناسه تا این لحظه.
✅ دریافت 260000 شناسه تا این لحظه.
✅ دریافت 270000 شناسه تا این لحظه.
✅ دریافت 280000 شناسه تا این لحظه.
✅ دریافت 290000 شناسه تا این لحظه.
✅ دریافت 300000 شناسه تا این لحظه.
✅ دریافت 310000 شناسه تا این لحظه.
✅ دریافت 320000 شناسه تا این لحظه.
✅ دریافت 330000 شناسه تا این لحظه.
✅ دریافت 340000 شناسه تا این لحظه.
✅ دریافت 350000 شناسه تا این لحظه.
✅ دریافت 360000 شناسه تا این لحظه.
✅ دریافت 370000 شناسه تا این لحظه.
✅ دریافت 380000 شناسه تا این لحظه.
✅ دریافت 390000 شناسه تا این لحظه.
✅ دریافت 400000 شناسه تا این لحظه.
✅ دریافت 410000 شناسه تا این لحظه.
✅ دریافت 420000 شناسه تا این لحظه.
✅ دریافت 430000 شناسه تا این لحظه.
✅ دریافت 440000 شناسه تا این لحظه.
✅ دریافت 450000 شناسه تا این لحظه.
✅ دریافت 460000 شناسه تا این لحظه.
✅ دریافت 470000 شناسه تا این لحظه.
✅ دریافت 480000 شناسه تا این لحظه.

پیشنهاد copilotاضافه کردن مکث 1 ثانیه ای به کد بود. تا سایت مارا بلاک نکند. اما من کد را تغییر دادم و برای محکم کاری مکث 5 ثانیه ای گذاشتم. ------> سه بار امتحان کردم و هربار روی همین عدد ایستاد.
راه حلی که به ذهنم رسید: دوبار کد بزنم و هربار 500000 ریکورد را استخراج کنم و درآخر باهم تلفیقشان کنم.
اینم نتیجه نداد و الان 14 ساعته که دارم تلاش میکنم و به نتیجه ای نمیرسم!

درنهایت:

دیتاست: توالی پروتئینی Cytochrome C در 2000 گونه جانوری.

توضیح کلی درمورد پروژه

برای جستجوی فازی (به این معنا که تنها تطابق کامل را با توالی موردنظر نشان نمیدهد و درصدهای مختلف را بررسی میکند) در توالی های پروتئینی FASTA، از سه کتابخانه استفاده کردم:

Streamlit: برای نمایش توابع در صفحه وب

Pandas: برای مدیریت داده ها و نمایش نهایی درقالب جدول

Fuzzywuzzy: برای پیاده کردن الگوریتم سریعی که درنهایت پیدا میکنم.

هدف من این است که یک توالی ورودی داشته باشم و توالی های مشابه را براساس قسمتی از آن پیدا کنم.

***نکته: من بدلیل محدودیت این برنامه ای که نوشتم 67 توالی موجود در فایل data.fasta استفاده کردم.**

منابع استفاده شده:

دوره ی آنلاین برنامه نویسی پایتون در سایت مکتبخانه

دوره پایتون برای زیست شناسان که ازطریق کانال دانشگاه با آن آشنا شدم

Copilot

Youtube

سایت فرادرس، سایت سون لرن و افراد متخصص در این زمینه برای راهنمایی های تخصصی تر و پیشرفت بهتر و متفاوت تر پروژه

پروژه دوم: داده تصویری

از آنجایی که کنجکاو بودم کار با داده تصویری چگونه است، درحد بسیار کلی تر نسبت به پروژه اصلی، به آن پرداختم و با راهنمایی و کمک گرفتن از علم و تجربه چندین نفر نتیجه ای به دست آمد.

- پروژه در محیط google colab اجرا شد
- نیازی به آپلود فایل نیست. تصاویر را از منبع آنلاین با کد Url وارد میکنیم.
- دیتا: سه تصویر از سایت wikimwdia:

Cell_structure.png

Cellular_mitosis.png

Dna_model.png

زهرا رحمانیان-40043313