Mode of Inheritance (MOI) Detection Task

شرکت پردیس ژن

مهدی حسن زاده

قسمت اول پروژه:

دانلود داده ها:

برای دانلود داده ها از سایت مخصوص هر پایگاه داده استفاده شده است. در طی این پروژه پایگاه داده های omim, cgd, clingen disease validity, gencc مورد بررسی قرار گرفتند. همچنین به دلیل پیدا نکردن داده مناسب در پایگاه داده gene2phenotype از بررسی آن صرف نظر شده است.

پایگاه داده omim:

در قسمت اول نوت بوک، فایل مربوط به پایگاه داده omim توسط کتابخانه pandas خوانده شده است. در ادامه با استفاده از تابع determine_inheritance برچسب های MOI از ستون phenotypes پایگاه داده استخراج شده است. تعداد برچسب های به دست آمده بدین صورت می باشد:

Unknown	756
AR	163
AD	114
AD/AR	34

برای تکمیل داده ها با استفاده از کتابخانه omim و تابع get_inheritance_mode به ازای هر سطری که برچسب MOI آن مشخص نیست از سایت پایگاه داده پرس جو می شود و در صورت پیدا شدن برچسب مناسب به داده ها اضافه می شود. تعداد اندکی برچسب (به طور مشخص ۲ عدد) به داده ها افزوده می شود.

تابع add_new_data:

با استفاده از این تابع بعد از خواندن پایگاه داده های دیگر، برچسب ها به داده ها اضافه می شود. ورودی این تابع شامل پایگاه داده قبلی، پایگاه داده جدید، نام ستون ژن در پایگاه داده جدید، نام ستون MOI در پایگاه داده جدید می باشد. این تابع خروجی ندارد و برچسب ها به پایگاه داده قبلی اضافه می شوند. با توجه به

درخواست پروژه داده های موجود در پایگاه داده اولویت بالاتری نسبت به داده های جدیدتر دارند. در نتیجه فقط سطر هایی که در پایگاه داده قبلی نیستند مورد بررسی قرار می گیرند.

پایگاه داده cgd:

ابتدا فایل مربوط به این پایگاه داده خوانده می شود. سپس برچسب های متفاوت موجود در این پایگاه داده نمایش داده شده است. سه مورد اول مورد توسط تابع add_new_data که در بالا ذکر شد استفاده خواهد شد. در نهایت توسط همین تابع داده های جدید افزوده می شوند. تعدادی برچسب جدید به داده ها اضافه می شوند که آمار آن در فایل نشان داده شده است.

پایگاه داده clingen:

ابتدا فایل مربوط به این پایگاه داده خوانده می شود. سپس برچسب های متفاوت موجود در این پایگاه داده نمایش داده شده است. دو مورد اول مورد توسط تابع add_new_data که در بالا ذکر شد استفاده خواهد شد. در نهایت توسط همین تابع داده های جدید افزوده می شوند. تعدادی برچسب جدید به داده ها اضافه می شوند که آمار آن در فایل نشان داده شده است.

پایگاه داده gencc:

ابتدا فایل مربوط به این پایگاه داده خوانده می شود. سپس برچسب های متفاوت موجود در این پایگاه داده نمایش داده شده است. با استفاده از تابع map برچسب ها به برچسب های قابل قبول برای تابع add_new_data که در بالا ذکر شد، تبدیل شده است. در نهایت توسط تابع add_new_data، داده های جدید افزوده می شوند. تعدادی برچسب جدید به داده ها اضافه می شوند که آمار آن در فایل نشان داده شده است.

داده های ابزار DOMINO:

داده های P(AD) که توسط ابزار domino تولید شده است را توسط کتابخانه pandas خوانده می شود. و در ادامه این داده ها با پایگاه داده اصلی توسط کتابخانه pandas ترکیب می شوند. آمار تعداد سطر های دارای P(AD) و سطر های خالی نشان داده شده اند.

قسمت دوم پروژه:

تابع generate_final_output:

این تابع پایگاه داده جمع شده در مرحله قبل و فایل vcf خوانده شده را به عنوان ورودی می گیرد. این تابع تا استفاده از مقدار pos موجود در فایل های vcf برچسب مناسب ژن مورد نظر را پیدا کرده و به همراه ستون های خواسته شده در توضیح پروژه، در پایگاه داده ای جدا ذخیره می کند. در صورتی که برای این pos نتیجه ای پیدا نشود مقدار None به ازای ستون های خواسته شده وارد می شود. در نهایت پایگاه داده جدید ایجاد شده برگردانده می شود.

خواندن فایل های vcf و انجام مراحل مورد نیاز:

در این قسمت بعد از مشخص کردن مسیر فایل های vcf، به ازای هر یک از فایل ها، تابع generate_final_output فراخوانی می شود و سپس خروجی آن در فایل مربوط به هر vcf ذخیره می شود. در ادامه تعداد برچسب های موجود در بعضی از فایل های vcf نشان داده شده است.

فایل های خروجی:

برای مشاهده نتیجه ها به پوشه outputs مراجعه بفرمایید.

منابع:

https://domino.iob.ch

https://omim.org

https://thegencc.org

https://research.nhgri.nih.gov

https://search.clinicalgenome.org

https://chat.openai.com