În cadrul acestui laborator am realizat o analiză integrativă multi-omics, scopul a fost explorarea relațiilor dintre datele genetice de tip SNP și datele de expresie genetică. Integrarea mai multor straturi omics permite o înțelegere mai profundă a mecanismelor biologice, evidențiind modul în care variațiile genetice pot influența nivelul de expresie al genelor.

Datele de intrare utilizate sunt in format CSV, iar analiza a fost realizată in Python folosind librarii specifice.

Am utilizat snp_matrix_MariusJalba.csv, matricea de SNP, care conține genotipuri pentru mai multe probe și expression_matrix_MariusJalba.csv, matricea de expresie genetică, cu niveluri de expresie de pentru aceleași probe.

Exercițiul 1 – ex01_PCA_and_viz.py

Am utilizat datele snp_matrix_MariusJalba.csv și expression_matrix_MariusJalba.csv

Am încărcat fișierele, aliniat probele apoi am normalizat prin z-score pentru a elimina influența diferențelor de scară între variabile. Am aplicat analiza PCA pe trei situații:

1. Pe datele SNP
2. Pe datele de expresie genetică
3. Pe datele combinate, obținute prin concatenarea celor două matrici

Pentru fiecare caz am generat imagini PNG, care ilustrează distribuția probelor în spațiul primelor două componente principale. Rezultatele PCA au evidențiat diferențe semnificative între structura datelor SNP și cea a expresiei genetice. Datele SNP prezintă o variație mai dispersată, specifică informațiilor genetice discrete, în timp ce datele de expresie genetică formează grupări mai coerente. PCA aplicată pe datele integrate reflectă contribuția ambelor straturi și permite observarea unor tipare complexe care nu sunt evidente în analizele separate.

Exercițiul 2 - ex02_cross_omics.py

În acest exercițiu am folosit matricea integrată multi-omics generată în exercițiul precedent, acesta conține SNP-uri și gene, organizate pe rânduri. Matricea integrată a fost separată în două subseturi. Pentru a evita rezultate eronate am eliminat liniile cu variație zero, apoi am calculat corelația Pearson între fiecare SNP și fiecare genă. Rezultatele au fost filtrate folosind un prag mai mare de 0.5, perechile SNP-genă au fost exportate într-un fișier CSV de ieșire, conform cerințelor. Ca rezultat analiza a identificat mai multe corelații semnificative între anumite SNP-uri și gene, sugerând relațiile funcționale între variațiile genetice și nivelurile de expresie genetică.

Concluzii:

Prin realizarea exercițiilor a fost demonstrată eficiența abordării integrative multi-omics în analiza datelor biologice complexe. PCA a permis explorarea vizuală a structurii datelor, iar analiza de corelație a evidențiat relații relevante între SNP-uri și gene.

Laboratorul subliniază importanța normalizării datelor, a integrării datelor, a integrării corecte a mai multor straturi omics și a utilizării metodelor statistice adecvate în bioinformatică. Rezultalele obținute oferă o bază solidă pentru studii avansate în domeniul analizei multi-omics.