Project title: Statistical Learning for Multi-Omics Data in Related Samples and Longitudinal Settings

Project reference: DT4H_10_2022

1st co-supervisor – Davide Pigoli – Department of Mathematics

1st co-supervisor –– Mario Falchi – Department of Twin Research & Genetic Epidemiology

Aim of the project

This project aims to develop and apply advanced statistical and machine learning methods to multi-omics data measured in related and unrelated subjects, and at different time points. Omics data are novel high-throughput molecular data from different layers of a biological system, whose integration helps deciphering the complex mechanisms underlying health and disease. However, integration methods that take into account the intra-subject and inter-subject effects and the time dependence are still in their infancy. The project will focus on the extension to the case of related samples and longitudinal data of existing methods for variable integration, such as Multi-Omics factor analysis, and for variable selection, such as Multi-Omics Sparse Partial Least Squares. The aim of the project is to develop and implement novel methods, test them on simulated data and then apply them to observational datasets, such as the TwinsUK dataset. The identification of genetic and clinical factors associated with the longitudinal variation of molecular data will help to better understand disease progression and the evolution of diseases risk, thus offering novel opportunities for prevention and intervention.

Project description

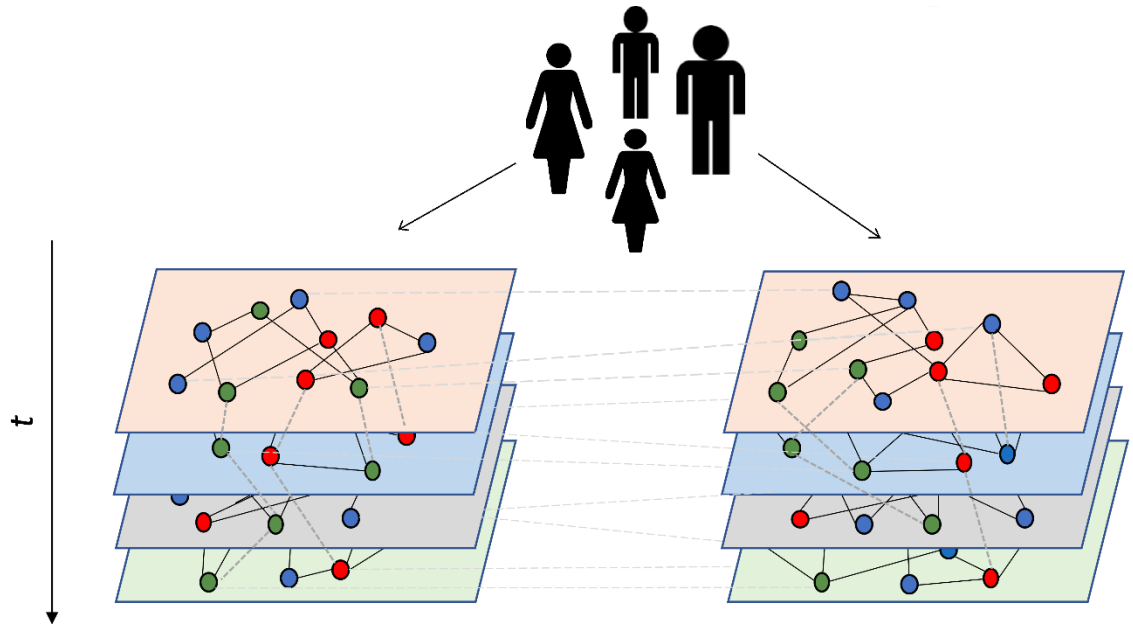
With the advance in high throughput technologies is now possible to characterise cohort studies for a wide range of molecular data in longitudinal settings, using prospective or retrospective biological samples. Together with clinical data and electronic health records, longitudinal multi-omics data creates new opportunities, but also several challenges. Indeed, if on one hand they allow a better understanding of physiology and aetiology, on the other hand statistical analyses gets more complicated, particularly because the nature of these data, characterised by multiple technical confounders, heterogeneous spacing between measurements, missing data, and sometimes non-independence between samples. This is the case for example of the TwinsUK cohort, where pairs of samples (twins) are genetically correlated, thus adding another level of complexity to the model. The TwinsUK cohort is a unique longitudinal collection of ~14,000 monozygous and dizygous twins with extensive clinical, lifestyle and nutritional data, and a plethora of multi-level molecular -omics data, including longitudinal genomic, epigenetic, metabolomic, glycomic, proteomic, and metagenomic data. Variable degree of relatedness between pairs of individuals is also typical in cohorts from isolated populations (e.g. Qatar Biobank), or in very large cohorts where multiple related individuals are likely to be sampled (e.g. UK Biobank or FinnGen).

Until now, most biomedical investigations have focussed on cross-sectional observational studies. The establishment or launch of large cohort studies in several regions worldwide, large-scale collaborative analyses, and availability of electronic health record data will shortly allow to investigate the dynamic evolution of an individual's health, to identify new early targets for disease prevention and progression.

Despite the surge of multi-omics studies there are not established methods and software for data integration in non-independent samples and longitudinal modelling, to fully exploits these multi-layered data. In this project, we will develop novel statistical methodologies to account for relationships between subjects and for the longitudinal structure of these data across the various steps of the analyses. For example, multi-omics integration based on sparse Principal Component Analysis, Multiple Factor Analysis, non-linear Multidimensional Scaling or Manifold Learning will be adapted to account for the dependence between repeated measurements, as well as the dependence between individuals. Variable selection and regression methods such as sparse Partial Least Squares Analysis, group LASSO, sparse Kernel methods and ensemble methods will similarly be extended to address these issues.

Finally, the project will apply these methods to the integration of multiple omic data in related samples and investigate the association between longitudinal trajectories and the co-variability of multiple traits with health and genetic data using data from TwinsUK and collaborators. Among these, the student will for example investigate what genetic variants are involved in the trajectory and variability of the metabolome and glycome and their effect on cardiometabolic traits, and the effect on systemic inflammation of the interplay between the gut microbial community and the host proteome and metabolome.

The student undertaking this project is expected to have a strong background in quantitative methods for high-dimensional data (penalised regression, mixed effects models, dimensional reduction). Previous experience with omics data is preferable but not essential.



Pictorial representation of observing multi-Omics data over time for a sample of individuals.