DeepIntegrOmics

END-TO-END DEEP LEARNING FOR PRECISION MEDICINE THROUGH METAGENOMICS AND COST-SENSITIVE DATA INTEGRATION

PRC ANR

CE45 - Mathématiques et sciences du numérique pour la biologie et la santé

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

anr agence nationale de la recherche
AU SERVICE DE LA SCIENCE

Labélisations

cap-digital Paris Region

medicen PARIS REGION innovation for health

UMMISCO Unité Mixte Internationale

SORBONNE UNIVERSITÉ

IRD Institut de Recherche pour le Développement FRANCE

nutriomics

MÉDECINE SORBONNE UNIVERSITÉ

Inserm

LAMSADE UMR 7243

Dauphine UNIVERSITÉ PARIS | PSL

cnrs

ibisc

université evry val-d'essonne

université PARIS-SACLAY

Project key information

Project leader : Jean-Daniel ZUCKER, UMMISCO, IRD, France;
Project duration : 42 months; **Starting date : February 2022** ; Ending Date: January 2026)
IRD budget : 277 K€; Total budget : 621 K
Keywords: Deep Learning; Predictive analytics; Large-scale machine learning and AI for life sciences; Bioinformatics; Metabolic diseases; Metagenomics

Partner institutions

IRD-SU / UMMISCO; SU-INSERM NUTRIOMCS; LISC-Université d'Evry, LAMSADE (University PSL)

Abstract

In chronic diseases such as cardiometabolic diseases (CMD), the use of intestinal microbiota as a source of patient stratification and of innovative treatment is on the rise. As a "super integrator" of the patient's condition metagenomics is poised to play a key role in precision medicine. However, there are still computational barriers to its routine use in medical services. In particular most metagenomics diagnosis approaches rely on tedious and computationally heavy projections of the sequence data against very large genomic reference catalogs (>170 Million genes for the latest one UHGP). Deep learning has revolutionized predictive analysis, improving many of the previous models involving heavy bioinformatics pipelines to perform classification or stratification tasks. Yet, very little literature exists on end-to-end deep learning of raw metagenomics data to stratify patients' cohorts and/or predict patient phenotypes.

Objectives

A first scientific barrier this project addresses is to develop metagenomics-based routine "point-of-care" prognosis or diagnosis. A recurrent problem in precision medicine is to integrate different sources of omics data, while controlling the cost/benefit balance of exams, in order to evaluate the usefulness of requesting more exams is critical to their routine use. Although CMD, in particular ischemic heart disease (IHD) and stroke, are the leading cause of global mortality and a major contributor to disability, current patient stratification is insufficient and integrated molecular signatures that inform on the evolution of CMD stages are missing. In this context the DeepIntegrOmics project main scientific goal is to significantly improve DL-based methodological frameworks using multi-Omics data for Precision Medicine in two main directions : first to support both reliable end-to-end prediction from metagenomics raw-data and second to improve classification accuracy and stratification by integrating other omics data. Two more applied objectives are to propose novel approaches for multi-omics biomarker identification of cardiometabolic disease stages and propose means of patient stratification through the interpretation of these neural network architectures.

Data and Deliverables

This study will be performed on a unique phenotypic database of 1844 patients (one of the largest existing datasets from the EU H2020 MetaCardis project) for which metagenomic, clinical and three types of metabolomic data are available. We will evaluate the classification performance of the DL integration architecture to predict the eight CMD groups (including control) to which the 1844 patients belong. We will assess the prognostic value of the stratification to predict CMD progression for 807 patients from the 1844 for whom we have characterized their evolution (clinical changes) during 10 years. Altogether, these objectives will support translational and precision medicine (i.e. classification and novel stratification of patients) in the perspective of deploying these models for routine use in clinical centers. From a translational perspective, the expected results in both stratification of patients in MetaCardis, biomarkers signatures and the ability to predict transition in disease progression are key outcomes that could help improve the management of patients with cardiometabolic diseases (CMD). From a methodological perspective, the expected result is both a DL architecture for cost-sensitive data integration and open sourced embeddings to perform multi- omics classification. In terms of impact, the classification based on the new gut microbiota-derived markers "omics" could generate new therapeutic targets. We also expect an Impact on patient management and the patients themselves.

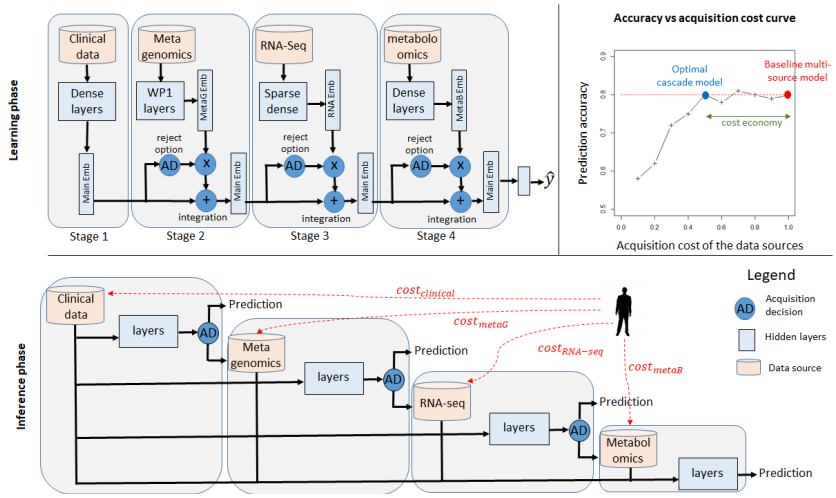

***Figure 2**: A cascade model from 4 data sources, both in learning and inference. Each blue rectangle represent a block of several hidden layers. The cascade model represents a trade-off between accuracy and acquisition costs.*

Impact and benefits of the project including applications in the south

Thanks to the federated database analyses and the proposed analytical work with partners having complementary expertise, DeepIntegrOmics will be able to propose new definitions of CMD based on meaningful characteristics of subgroups, according to the stratification approach. This will provide a more holistic view of patients with CMD encouraging a shift in thinking when it comes to this complex disorder. We also expect an Impact on patient management and the patients themselves. A major result of this project will be a competitive pre-trained contextualized embedding model that delivers a significant performance boost for real-world disease-prediction problems as compared to state-of-the-art models learning from quantitative metagenomics data. The novelty of such an approach is to be based on raw reads and to support end-to-end classifications of disease. The second result will be a reusable fully instantiated Deep Learning architecture integrating embeddings for different omics types of data to stratify and classify patients.
A final objective for UMMISCO is to democratize the technologies developed in the deepIntegromics project, and involve several students from the International Doctoral.

References

Prifti, Edi, Yann Chevaleyre, Blaise Hanczar, Eugeni Belda, Antoine Danchin, Karine Clément, and Jean-Daniel Zucker. 'Interpretable and Accurate Prediction Models for Metagenomics Data'. *GigaScience* 9, N°3 (2020): doi:10.1093/gigascience/giaa010.

Hanczar, B. & Bar-Hen, A. Controlling the Cost of Prediction in using a Cascade of Reject Classifiers for Personalized Medicine. Proc 9th Int Jt Conf Biomed Eng Syst Technologies 42–50 (2016) doi:10.5220/0005685500420050.