Mini-symposium JOBIM 2024

Machine Learning and Statistics in Genomics and Metagenomics

Apprentissage automatique et statistiques en génomique et méta-génomique

The mini-symposium is jointly organised by the LEGO and StatOmique working groups, and supported by the CNRS GDR BIMMM and the IUF.

Organisateurs:
Julie Aubert, Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, Paris.

Marie-Agnès Dillies, Institut Pasteur, PF2 Plateforme Transcriptome et Epigénome, Paris.

Christelle Hennequet-Antier, Université Paris-Saclay, INRAE, BioinfOmics, MIGALE Bioinformatics Facility, Jouy-en-Josas.

Laurent Jacob, Sorbonne Université, CNRS, IBPS, Laboratory of Computational and Quantitative Biology (LCQB), UMR 7238, Paris.

Flora Jay, Université Paris-Saclay, CNRS, INRIA, LISN, Paris.

Elodie Laine, Sorbonne Université, CNRS, IBPS, Laboratory of Computational and Quantitative Biology (LCQB), UMR 7238, Paris.

Raphaël Mourad, MIAT, INRAE, 31320, Castanet-Tolosan, University of Toulouse, UPS, 31062, Toulouse.

Actuellement, une initiative mondiale est en cours pour séquencer la biodiversité de la Terre, produisant des volumes de données génomiques et métagénomiques sans précédent. Dans ce symposium, notre objectif est d'explorer les récentes avancées méthodologiques en matière d'apprentissage automatique et de statistiques pour exploiter ces données. L'accent sera mis sur l'amélioration de notre compréhension de la relation génotype-phénotype et du rôle du microbiote dans la santé et les maladies humaines. Nous aborderons différents défis (variabilité inter et intra-population/espèces, intégration de données omiques compositionnelles) et nous nous pencherons sur le potentiel et les limites des grands modèles de langage pour les séquences biologiques.

16h30-17h00

Mafalda Dias, Center for Genomic Regulation (Spain)

Modelling the genetic variation across the tree of life to learn about human disease

The genetic variation observed across the tree of life is the result of millions of years of

evolutionary experiments and should therefore contain valuable information to link genotype to molecular function and ultimately to the genetic architecture of disease. Deep generative modelling is a powerful approach for describing the rich distribution of observed variation, and revealing the patterns of constraint in sequence space that must be preserved in order to maintain fitness. In this talk I will describe our recent progress in developing models to aid in genetic diagnosis and disease gene discovery. We propose a model which places variants on a proteome-wide scale of pathogenicity and show that in contrast to previous models, this now enables us to identify causal variants from whole exome data in rare disease patients. Using this model, we find evidence for over 100 novel genetic disorders.

17h00-17h30

Jean-Daniel Zucker, Sorbonne Université, IRD, UMMISCO (France)

Challenges of Learning Embeddings from (Raw) Metagenomics Data

The field of metagenomics offers a panoramic view of microbial communities' genetic material, presenting both opportunities and challenges for computational biology. A primary hurdle in leveraging this data is the development of robust, informative embeddings that can capture the complex, high-dimensional relationships inherent in metagenomic samples. This is one of the objectives of the DeepIntegrOmics ANR project. Traditional machine learning approaches often struggle with the multiple-instance nature of metagenomic data, leading to embeddings that incompletely represent the underlying biological phenomena. This talk will discuss existing embeddings of Metagenomics data and explore how meta-learning can be an approach to the problem of finding good embeddings.

17h30-18h00

Baptiste Ruiz, IRISA, Université de Rennes, INRIA, CNRS (France)

SPARTA: a knowledge integration-based pipeline for disease state classification through the robust selection of inter-associated OTUs and functions

The field of personnalized medicine has major stakes in using an individual’s microbiota as a descriptor of health. This raises the question of the interpretability of microbiotal signatures found for various diseases. To gain insight on this matter, we developed the SPARTA (Shifting Paradigms to Annotation Representation from Taxonomy to identify Archetypes) pipeline to highlight and interlink significantly discriminating Operational Taxonomic Units (OTUs) and metabolic functions. SPARTA relies on the integration of the information from the UniProt database concerning the gut microbiota’s functional annotation to OTU abundance data, and on Machine Learning classification. Iteration of this method can shrink the list of the microbiotas’ descriptors tenfold, both in terms of OTUs and metabolic functions. It also reveals that the shift to functional profiles comes at no significant cost to overall classification performance. Finally, we highlight how discriminant metabolic functions may arise from the aggregation of several low-abundance OTUs, giving visibility to these functions which are therefore not easily derivable from OTU-based approaches, marking them as potentially novel leads.

18h00-18h30

M.Luz Calle, University of Vic - University Central de Catalunya (Spain)

coda4microbiome: predictive modeling with compositional covariates in microbiome studies

M.Luz Calle[1], *Meritxell Pujolassos*[1], *Antoni Susín*[2]

[1]Bioscience Department, Faculty of Sciences, Technology and Engineering, University of Vic – Central University of Catalunya, Vic, Spain

[2]Mathematical Department, UPC-Barcelona Tech, Barcelona, Spain

Prediction models play an important role in establishing the relationship between a set of covariates and an outcome of interest. Selecting the appropriate variables that are included in the model is often one of the most important and difficult parts of model building. This is even more challenging when the set of covariates form a composition and the goal is to identify which components are more associated with the outcome. This situation arises in many fields, for instance in microbiome studies where the interest is to identify which microorganisms are associated to a disease.

Microbiome data is compositional since raw abundances and the total number of sequences for each sample are not by itself informative, as they depend on technical issues such as laboratory sample preparation and sequencing depth. The total sum constraint of microbiome relative abundances induces important dependencies between the components.

In this talk I will discuss the effects of ignoring the compositional structure of microbiome data and will present *coda4microbiome*, a new methodology for analyzing microbiome data within the Compositional Data Analysis (CoDA) framework. The algorithm relies on the analysis of log-ratios between pairs of components and variable selection is addressed through penalized regression on the “all-pairs log-ratio model”, the model containing all possible pairwise log-ratios. The algorithm is implemented for cross-sectional, longitudinal and survival studies.