SMPGD 2024 Paris
Abstracts

Thursday 1 February

10:10 - 11:05 20 years of phylogeny

Céline Scornavacca, Institut des Sciences de l'Evolution, Montpellier, Fr
20 years of methodological developements in phylogeny: phylogenetics, phylogenomics, phylodynamics, and much more!

11:05 - 12:05 Contributed session

Andreoletti Jeremy, Institut de Biologie de l'ENS (IBENS)
Incorporating Fossil Occurrences into Birth-Death-Fossilization Models through Bayesian Data Augmentation
Traditional phylodynamic models in macroevolutionary biology often overlook the wealth of information available in fossil occurrences that cannot be directly integrated into the phylogenetic tree. Addressing this gap, this work focuses on the assimilation of such occurrences into birth-death-fossilization modeling frameworks, which not only leverages the entire fossil record but also enables the application of these models to epidemiological studies involving prevalence data. To tackle computational challenges, we employ Bayesian Data Augmentation methods, making it feasible to compute likelihoods for even highly complex models. Additionally, the work includes an empirical analysis of cetacean diversification, investigating the relative influence of various types of data on model outputs, and the impact of model misspecification on inference quality.

Goedhart Jeroen, Amsterdam University Medical Centers – Netherlands
Co-data Learning for Bayesian Additive Regression Trees
One of the promises of omics data is to improve the diagnosis of cancer and to find relevant biomarkers that may be used for therapy. However, omics data is typically high-dimensional, which, combined with the complicated interaction patterns between the measured omics covariates, poses significant challenges for prediction and feature selection. To improve prediction and feature selection, we propose to incorporate co-data, i.e. external information on the measured covariates, into Bayesian additive regression trees (BART) [1], a sum-of-trees prediction model that utilizes priors on the tree parameters to prevent overfitting. BARTs ability to model nonlinearities combined with co-data to guide the search for relevant variables may serve as an interesting tool for omics-based prediction models.
To incorporate the co-data, we develop an Empirical Bayes (EB) framework that estimates, assisted by co-data, the hyperparameters which determine prior covariate weights in the BART model. Our proposed method can handle multiple types and sources of co-data, whereas most existing methods only allow co-data in the form of groups [2,3]. Furthermore, our proposed EB framework enables the estimation of the other hyperparameters of BART as well. Hyperparameters of BART are typically estimated using cross-validation.

Empirical Bayes avoids using an arbitrary grid and may therefore render more refined hyperparameter estimates. We show that our method renders both improved predictions and variable selection compared to default BART in simulations. Moreover, it enhances prediction and variable selection stability in an application to diffuse large B-cell lymphoma diagnosis based on mutations, translocations, and DNA copy number data. Furthermore, our method is competitive to state-of-the-art co-data learners such as ecpc [4] and corf [5].

Fallet Sara, Univ. Bordeaux, INSERM, INRIA

Conditional cumulative distribution function testing for gene set analysis of single-cell RNA-seq data

Single-cell RNA-seq (scRNA-seq) technology measures gene expression in hundreds or even thousands of cells from a single biological sample, allowing to study molecular mechanisms at the single-cell resolution. In immunology, this technology is increasingly used to disentangle the complex immune response to infection (or vaccination) while accounting for cellular heterogeneity in, the blood. Differential Expression Analysis (DEA) allows to identify which genes are differentially expressed across different conditions, cell types, timings or exposures. However, DEAs often encounter challenges related to statistical power and stability, notably due to the dynamic nature of gene expression and cellular state heterogeneity. Investigating instead gene sets associated with specific immune functions, derived from prior biological knowledge, can enhance the statistical power and stability of the analysis while facilitating the biological interpretation of results.

We introduce a novel gene set analysis method tailored for scRNA-seq data. This method relies on the estimation and testing of conditional distribution functions, eliminating the need for distributional assumptions. This new method is suitable for complex experimental designs, testing the association of each gene set with one or multiple variables of interest (whether continuous or discrete), while potentially adjusting for additional covariates. We apply this new methodology to two single-cell RNA-seq real dataset investigating the immune response to SARS-CoV-2 infection in humans, with respectively 84,140 virus-reactive CD8+ T cells from 38 patients [1]; and 1,191,463 peripheral blood mononuclear cells from 222 donors [2].

Kusnadi et al. (2021). Severely ill COVID-19 patients display impaired exhaustion features in SARS-CoV-2-reactive CD8+ T cells. Science immunology 6(55).

Aquino et al. (2023). Dissecting human population variation in single-cell responses to SARS-CoV-2. Nature 621(120-128).

Van de Voorde Michael, Department of Plant Biotechnology and Bioinformatics, Ghent University

Single-plant omics: Profiling individual plants in a field to identify processes affecting yield

In plant science, translating knowledge from the lab to the field is often not straightforward because field conditions are very different from controlled lab conditions. Studying plants directly in the field with controlled experiments can be costly and the level of control that can be achieved is often limited. Observational studies with uncontrolled perturbations (e.g. environmental) on the other hand can be more easily set up in a field. Observational data however come with their own array of challenges; plants in realistic field conditions are bombarded by a myriad of different interacting stressors, both biotic and abiotic, and many of the perturbations influencing the study subjects may remain unobserved and hence unknown. It is generally also much more challenging to establish cause-effect relationships from observational data alone. Nevertheless, even merely correlational data generated in the field may help narrowing down the lab-field gap in plant science. In this perspective, we are developing a single-plant omics strategy to study the molecular wiring of plant traits directly in the field, based on profiling of individual field-grown plants: during a recent field trial, we profiled the autumnal rosette leaf transcriptome and a range of phenotypes (both before winter and at time of harvest) of 192 plants of winter-type rapeseed variety Darmor, along with several environmental data layers at individual plant resolution such as microbiomes, soil nutrient profiles and measurements from environmental sensors. To analyse this extensive spatial multi-omics dataset we implemented factor analysis models for integration of spatial omics data (e.g. MEFISTO), linear mixed-effects models that take into account the spatial autocorrelation structure of the data, elastic net models and random forest regression. The latter are implemented to model plant phenotypes as function of other data layers such as autumnal gene expression, nutrients and microbiomes, and help in identifying features that potentially influence plant yield: we found *important features in our yield models to include genes involved in vegetative to reproductive phase transition and floral transition, indicating that developmental processes in autumn influence final yield in summer. Conceptual similarity between single-plant and single-cell data inspired us to apply methods from the single-cell field such as trajectory inference on our single-plant data to further unravel these developmental effects: transcriptome data at plant level allows us to order plants along a developmental trajectory in the same way as single-cells can be ordered along a trajectory of differentiation. An inferred pseudotime value for each individual plant corresponding to their (pseudo-)developmental state can then be correlated to the measured environmental variables to gain more insight into how variability in plant development before winter, and its impact on final plant yield, can be explained by the environment at plant level.*

13:45 - 14:40 20 years of optimization

Francis Bach, Laboratoire d'Informatique de l'École Normale Superiéure, CNRS/ENS/INRIA, Paris, FR

Mixing statistics with optimization: from stochastic gradient decent to double descent

In this talk, I will present an overview of the long joint history of the interface between statistics and optimization, showing how they influence each other and can challenge their core principles. In particular I will focus on the impact of gradient-descent techniques for large over-parameterized models.

14:40 - 17:15 RNA structure and function

Camille Marchet, CNRS, Lille, FR

Reference-free transcriptomics and other large indexes

The field of eukaryote pangenomics has predominantly utilized variation graphs to understand genomic diversity and structure. However, a wealth of alternative approaches giving access to collections of genomes or transcriptomes exists, within less accessible scientific literature, often overlooked by the bioinformatics community. In this talk, I will present an overview of these lesser-known techniques, highlighting their properties and potential advantages over conventional methods. I will discuss the current promises and limitations of these approaches. Additionally, I will review various applications of these novel structures, notably in transcriptomics.

Nikolay Shirokikh, John Curtin School of Medical Research, Canberra, AU

Comprehensive translational profiling and STE AI to measure absolute protein biosynthesis rates and rapid changes in mRNA usage

Translational control is important in all life but remains a challenge to accurately quantify. When ribosomes translate messenger (m)RNA into proteins, they attach to the mRNA in series, forming poly(ribo)somes, in which ribosomes can co-localise. By analysing ribosomal co-localisation on mRNA using enhanced translation complex profile sequencing (TCP-seq) based on rapid in vivo crosslinking, we detect long disome footprints outside areas of non-random elongation stalls. We further show that these footprints are linked to translation initiation and protein bioproduction rates, providing a previously-missed feature of functional significance.

Applying advanced machine learning to the comprehensive ribosome localisation patterns we derive from the rich in vivo footprinting data for the first time, we create a novel, accurate and self-normalised measure of translation (stochastic translation efficiency, STE). STE has new applications in interrogating mRNA function and performance in live cells, and dissecting cell states in disease pathophysiology and drug development. Using STE to study nutrient starvation in yeast as a model system with extremely fast response, we refine translational control from other rapid RNA changes, and highlight metabolic rearrangements invoked by the cells solely at the translational level. In this system, we show that STE is invaluable for identifying the ab-

solute translational ranking of mRNA and its control elements under specific conditions. We envisage STE will aid the development of next-generation synthetic biology designs and mRNA-based therapeutics.

Emiliano Ricci, ENS Lyon, FR

Epitranscriptomics : a dynamic landscape of RNA modification marks to orchestrate cell function

RNA is a versatile molecule that can carry genetic information but also catalyze enzymatic reactions that are essential to life. From non-coding to messenger RNAs, a myriad of RNA species undergoes post-transcriptional chemical modifications, influencing secondary structures, stability, and interactions with cellular proteins. This intricate epitranscriptomic code is both constitutive, integral to fundamental RNA functions, and dynamic, actively participating in the fine-tuning of gene expression.

In this presentation, we will describe the principal types of post-transcriptional RNA modifications prevalent in eukaryotic cells, exploring their functional roles from mRNA translational control to contributions in innate immunity. Furthermore, we will review the different technical approaches that exist to map and quantify RNA modification sites. Lastly, we will explore the potential therapeutic and technological applications of harnessing RNA modifications, highlighting the promising avenues for leveraging this epitranscriptomic knowledge.

Virginie Marcel, Cancer Research Center of Lyon, FR

Alteration of ribosomal RNA 2'O-ribose methylation in cancer: a novel epitranscriptomic mark involved in translational regulation

While for over 40 years the ribosome was considered as a neutral player in gene expression, it is now emerging as a key regulator of translation through modulation of its composition, notably via chemical modifications of ribosomal RNAs (rRNAs). Among the various chemical modifications of rRNA, the 2'O-ribose methylation (2'Ome) is the most abundant and is present at about 110 specific nucleotide positions. Using cellular models, we revealed that rRNA 2'Ome can be a source of ribosome diversity. We report that 2'Ome is altered during tumorigenesis and directly affects the translational behavior of ribosomes toward mRNA subsets (1,2). To determine its alteration in tumor biopsies, we used an innovative omic technology, the RiboMethSeq dedicated to rRNA 2'Ome profiling (3), for which we contribute to develop methodologies and bioinformatic tools (https://github.com/RibosomeCRCL). Using RiboMethSeq, we demonstrated that rRNA 2'Ome is altered in different types of cancer, including breast cancer and glioma (4,5). Furthermore, our data showed that variability in rRNA 2'Ome levels is restricted to particular rRNA sites and associated with different clinical features. Our data support the recent entry of the ribosome into the emerging field of epitranscriptomics (6).

17:15 - 17:45 Contributed Session

Liehrmann Arnaud, Institut des Sciences des Plantes de Paris-Saclay

DiffSegR: An RNA-Seq data driven method for differential expression analysis using changepoint detection

To fully understand gene regulation, it is necessary to have a thorough understanding of both the transcriptome and the enzymatic and RNA-binding activities that shape it. While many RNA-Seq-based tools have been developed to analyze the transcriptome, most only consider the abundance of sequencing reads along annotated patterns (such as genes). These annotations are typically incomplete, leading to errors in the differential expression analysis. To address this issue, we present DiffSegR - an R package that enables the discovery of transcriptome-wide expression differences between two biological conditions using RNA-Seq data. DiffSegR does not require prior annotation and uses a multiple changepoints detection algorithm to identify the boundaries of differentially expressed regions in the per-base log2 fold change. In a few minutes of computation, DiffSegR could rightfully predict the role of chloroplast ribonuclease Mini-III in rRNA maturation and chloroplast ribonuclease PNPase in (3'/5')-degradation of rRNA, mRNA, and tRNA precursors as well as intron accumulation. We believe DiffSegR will benefit biologists working on transcriptomics as it allows access to information from a layer of the transcriptome overlooked by the classical differential expression analysis pipelines widely used today. DiffSegR is available at https://aliehrmann.github.io/DiffSegR/index.html.

Richard Hugues, Robert Koch Institute

Assessing conservation of alternative splicing with evolutionary splicing graphs

Alternative splicing (AS) can significantly expand the proteome in eukaryotes by producing several transcript isoforms from the same gene. AS has been linked to morphological diversity, organ development, disease susceptibility, immune adaptation and interactome rewiring, among others. Although AS is well described at the genomic level, little is known about its contribution to protein evolution and the extent of the contribution of AS to proteome diversity has been a matter of debate. This question could be accurately addressed using evolutionary conservation. There is a clear need for computational methods that can couple the diversity of proteoforms resultings from AS with measures of sequence conservation across species.

Our work introduces a new method enabling for the first time granular estimates of alternative splicing conservation. It significantly improves our knowledge about the amount of functionally relevant variations. We first determine orthology relationships between exonic regions in the context of alternative splicing by extending Multiple Sequence Alignments to Splicing Graphs structure. We construct an Evolutionary Splicing Graph (ESG) where nodes define orthologous exon groups (denoted s-exons) and paths in the graph correspond to transcripts [2]. The ESG summarizes the transcript variability observed across species, allowing the direct detection of conserved alternative splicing events. By analyzing AS conservation as far as teleosts, we show a clear link between the functional relevance, tissue-regulation and conservation of alternative splicing events on a set of 50 human genes. By constructing ESGs for the whole human proteome, we could annotate 46,000 evolutionary conserved AS events coming from 8,000 human protein-coding genes, dramatically changing previous estimates [2]. We further identified a few thousands of genes where alternative splicing modulates the number and composition of pseudo-repeats shared across species, thus demonstrating the widespread alternative usage of protein repeats in modulating protein interactions and opening avenues for targeting repeat-mediated interactions. [2,4].

The set of orthologous s-exons can be further analyzed to infer evolutionary scenarios explaining the observed transcripts variability. Each transcript history is described by a tree encapsulated in the tree of species, resulting in a phylogenetic forest [1]

Both works are made accessible as open source tools to the community. A web server (http://www.lcqb.upmc.fr/Ases), integrating both tools and providing interactive output is available [3]. It facilitates the study of alternative splicing evolution and its relation with the observed protein diversity by enabling interactive and linked representations of the evolutionary splicing graph and phylogenetic forest that were constructed.

Friday 2 February

9:00 - 9:55 Digital Twins

Loïc Paulevé Laboratoire Bordelais de Recherche en Informatique, CNRS Bordeaux, FR

Synthesis of executable models of cellular dynamics from knowledge and experimental data

Gautier Stoll, Laboratoire Immunologie et Cancérologie Intégratives, INSERM / Univ. Paris Descartes, Paris, FR

Mathematical model of CAR T cell therapy

CAR T cells are T lymphocytes that have been engineered with a "Chimeric Antigen Receptor" for targeting specifically cancer cells. They are constructed from individual patient T lymphocytes and are reinjected as treatment. For some cancers, for instance in multiple myeloma, they represent a new and promising therapy. An important clinical challenge is the understanding and the optimization of such a treatment.

We will present a mathematical model (or digital twin) of CAR T cell action on tumor cell. The model is based on signaling pathways inside T cell and tumor cells, with their interactions. The mathematical approach is based on Boolean modeling, where gene/proteins are either active or inactive. For that, we integrate our model in UPMaBoSS tool, which is perfectly suited for time-dependent interactions between heterogeneous cell types, including their respective signaling pathways.

We will show the different strategies for using the model as a virtual human cohort. As preliminary results, we will show how we can integrate clinical data regarding cell type abundance.

Anna Niarakis, Centre de Biologie Integrative, Univ. Toulouse III - Paul Sabatier, and Lifeware, INRIA-Saclay, FR

Building immune digital twins for complex human pathologies - a community effort

Jieling Zhao, SIMBIOTX, INRIA, Palaiseau, FR

Towards a full digital liver twin of drug-induced liver injury, fibrosis and disease progression.

Digital twin is the projection of the real world object to the digital world. It has been widely used to study the biological and medical phenomena such as to understand the biomechanical growth control mechanisms of liver regeneration and to explore the extrapolation strategies for drug-induced liver injury. In this talk, we present a digital twin of the liver and its application to drug-induced liver damage, liver regeneration and fibrosis formation as a prominent example of a disease process. This digital twin is based on a biophysics-based computational model which can accurately capture the deformation of cells, capillaries, and extracellular matrix according to their biomechanical properties. As drug-induced damage overdosing paracetamol (acetaminophen) is studied in a multilevel model integrating drug detoxification in each individual hepatocytes according to the processes in the respective liver zones (zonatation). The regeneration process triggered by the drug is based on a complex cross-talk between cells exchanging extracellular signals. This intercellular signaling network is integrated into the digital twin to allow the communication between various cell types through corresponding signals. We show that for the application of liver regeneration, the digital twin could help identify a set of successful alternative mechanisms controversly discussed in the biological and medical community for a perfect liver recovery and predict the effect of depletion certain cell types. Repetitive damage has been shown to cause fibrosis characterized by deposition of extracellular matrix but the mechanism leading to the characteristic spatial pattern of fibrosis are not understood. The digital twin proposes a mechanism of how the fibrotic pattern is formed. In summary, here we show the potential of the digital twin for studying complex biological/medical problems at subcellular level and its role as a pillar complementary to real-world experiments in future.

11:35 - 12:05 Contributed Session

Hawinkel Stijn,Department of Plant Biotechnology and Bioinformatics, Ghent University

Probabilistic indices for a simplified analysis of point patterns from spatial omics technologies

Emerging spatial omics technologies allow the study of biology in it spatial context. Techniques with single-molecule resolution necessitate custom analysis, as there is no quantitative readout: only locations of the different molecules are measured. Yet these point patterns need to be analysed jointly despite them having varying shapes and dimensions, which complicates warping (overlay). Current analysis are often ad-hoc and prone to visual cherry picking.

As a remedy, we developed the spatrans methodology, which allows to combine multiple point patterns with multiple levels of nesting in a single analysis. We use probabilistic indices as outcome values, which are unitless and therefore comparable across point patterns of varying shapes and dimensions. We argue that they are simpler and more interpretable than the usual analyses based on Ripley's K-function. The variance of the outcome is estimated by information sharing across all features. Our method allows to test for aggregation close to the cell boundary or center, as well as for aggregation of genes and coexpression of gene pairs. In addition, we can test whether these colocalization patterns depend on a condition. We argue that the classical null hypothesis of complete spatial randomness is not always the most informative ones, and testing should proceed conditional on the structure of the tissue under study.

We illustrate our method on datasets from mouse, bacterial biofilm and plant root and contrast the outcome with the author's original results. Our different conclusions underscore the danger of purely visual inspection of point patterns, and emphasises the need for reproducible statistical analysis.

Dick Nicola, Centro de Regulación Genómica

TAFI (Tumor Allele Frequency Interpreter): a new deep learning tool to reveal the evolutionary history of tumors

Tumor progression is a somatic evolutionary process in which population expansion is driven by the accumulation of mutations that promote cell proliferation (cancer drivers). Other mutations, called passengers, with an indifferent fitness effect, subsequently accumulate. The distribution of both driver and passenger variant allele frequencies (VAF) can be used to infer relevant biological parameters, such as mutation rates, or to distinguish between demographic models. However, some technical issues arise when trying to compute these estimators from available data. In large scale sequencing projects (e.g., PCAWG), the technique used is bulk sequencing, with low read depth. This curtails our ability to call low frequency variants and yields a general underestimation of the amount of genetic diversity in the tumor. Our goal is to develop an algorithm that can estimate mutation rate and demographic history while accounting for these systematic errors and biases generated by sequencing techniques, mutation calling tools and subsequent filters, that are applied to generate current datasets. Our method combines simulations and deep learning, to estimate the true amount of genetic diversity and provide reliable estimates of the tumor's mutation rate, demography and age. This algorithm could be applied to create a comprehensive study of how different evolutionary parameters vary across tumor types and across individual tumors.

13:45 - 14:40 10 years of Single-Cell

Céline Vallot, Institut Curie, Paris, FR

10 years of single-cell research, applications in cancer epigenomics

Accessing the identity and function of single cells within complex tissues has been a long-standing quest in biology. That heterogeneity exists in biological systems such as cell populations, organs and tumors is widely accepted in modern biology, yet measuring this heterogeneity in an exhaustive fashion without any a priori expectations has only recently become possible thanks to single cell genomics. It all started in 2015 with the advent of RNA barcoding, the ability to add a unique, cell-specific genetic label (a barcode) into all RNA molecules of a cell, with thousands of cells processed in parallel droplets. The scope of possible *applications of this method was immediately apparent and rapidly spurred the creativity within the scientific community. Any RNA or DNA species of interest can be profiled at single-cell resolution by simply barcoding it in droplets, before releasing it and sequencing it. A plethora of single-cell omics methods flourished for the genomic, transcriptomic or epigenomic profiling of complex biological systems. We will discuss the advent of single cell genomics and its application to study epigenomic heterogeneity. In addition, we will showcase its application to breast cancer.*

14:40 - 14:55 Contributed Session

Chaussard Alexandre, Sorbonne Universite, CNRS

Taxa-PLN: A Latent model approach for microbial interactions inference in the gut microbiome using taxa-abundance data

The gut microbiota is a complex ecosystem composed mostly of bacteria interacting with each other and their environment. If the composition of the microbiota has proven to be a relevant biomarker for several diseases, the specific structure of gut microbiota has poorly been taken into account. Indeed, the microbial composition is described by discrete and sparse (many null values) data, which also have a taxonomic structure. More precisely, each bacterium belongs to several groups that are hierarchically ordered from more precise (species) to less precise (domain). Although this taxonomic information is known, it remains unclear how it is related to the impact of a bacterium on its host. While some recent works (Chiquet et al. 2019) proposed a framework that accounts for the sparse and discrete nature of the microbiota, the impact of the taxonomic structure has not been investigated yet. In this work, we aim to introduce a new framework and dedicated algorithms that account for the sparse taxonomic abundance nature of the microbiota data. Our focus is to propose interpretable methods to unveil the complex interplay among bacterial species, with an application in the context of inflammatory bowel diseases.

The proposed approach, based on Poisson Log-Normal models, accounts for Markov dependencies to include the taxonomic tree structure linking the bacteria while enabling taxa of different branches to influence each other. This not only extends the scope of interaction modeling but also aligns with the inherent complexity and diversity of microbial communities. Additionally, we present a novel variational approach that incorporates the Markov structure of the posterior distribution to enhance the precision of the variational estimation. Such variational families allow us to obtain theoretical guarantees on latent state estimation and data reconstruction.

The generative structure of the model enables us to challenge our proposed framework on artificial data. The performance is assessed using typical microbial metrics like alpha and beta diversity as well as standard statistical indicators like graph dissimilarities. Our analysis explores the model characteristics, highlighting its capacity to faithfully replicate underlying microbial dynamics within the gut microbiome. To assess real-world applications, we conduct an empirical investigation on a cohort of patients diagnosed with Crohn's disease. The model is systematically benchmarked against state-of-the-art algorithms, providing a comprehensive analysis of its efficiency in capturing complex microbial interactions associated with pathological conditions.

15:20 - 16:15 20 years of latent models

Stéphane Robin, Sorbonne Université, Paris, FR

A partial history of latent variable models in genomics

Latent variable models have been involved, explicitly or implicitly, in the development of genomics since its very beginnings. Mixture models, stochastic Markov models or stochastic block models are just a few examples of latent variable models that have been used, for example, for gene detection, transcriptome analysis or protein interaction network analysis.

These models all assume that only some of the variables involved in the process under study are actually *observed. Inferring the parameters of such a model therefore poses specific problems, since it requires reconstituting part of the information concerning unobserved, or "latent", variables. The most common strategy, based on the EM algorithm, requires, for example, the evaluation of the conditional moments of the latent variables conditional on the observed variables. However, this approach proves impractical for models of even moderate complexity.*

We will present a series of applications of latent variable models in genomics and the statistical inference problems raised by each of them. In particular, we will show how the development of extensions to the EM algorithm have accompanied the increasing complexity of models used in genomics.